万卷方法

# 定量研究基础：测量篇

MEASUREMENT, DESIGN, AND ANALYSIS: AN INTEGRATED APPROACH

埃拉扎尔·J.佩达泽 (Elazar J.Pedhazur)
丽奥拉·佩达泽·施梅尔金 (Liora Pedhazur Schmelkin) 著

夏传玲 译

张彦 赵联飞 夏传玲 校



重庆大学出版社
http://www.cqup.com.cn

我们并不是说，无论专业和理论兴趣何在，所有专业人员一定要成为测量、设计和分析上的“专家”。但我们的确主张，这些研究领域以及它们之间的相互影响，对它们的一个基本理解，是成为研究结果的一个精明消费者的必要条件；如果我们想要成为一个胜任工作的研究者，更是如此。因此，本书的目标是帮助大家精通研究的各个方面，并帮助大家培养一种视角，让大家能够关照它们之间的相互关联和相互依赖。而且，我们希望帮助大家了解理论在研究工作中至关重要的指导作用。

——作者语

**发表及参阅相关讨论，请登录：**

· 万卷方法与学术规范博客圈（http//q.blog.sina.com.cn/fafang）



· 万卷方法与学术规范微博（http://weibo.com/cqupwjff）

ISBN 978-7-5624-7231-5



9 787562 472315 >
定价：25.00元

万卷方法 | 社会科学研究方法经典译丛



# 定量研究基础：测量篇

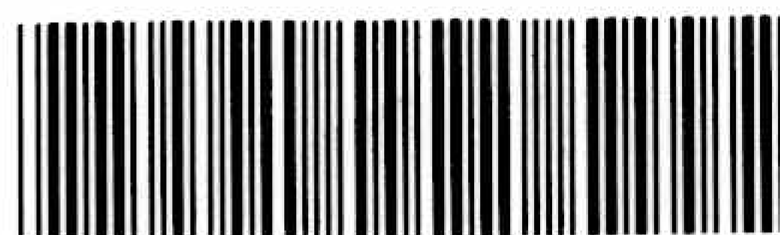
MEASUREMENT, DESIGN, AND ANALYSIS:
AN INTEGRATED APPROACH

埃拉扎尔·J.佩达泽 (Elazar J.Pedhazur)
丽奥拉·佩达泽·施梅尔金 (Liora Pedhazur Schmelkin) 著

夏传玲 译

张彦 赵联飞 夏传玲 校







重庆大学出版社

**图书在版编目(CIP)数据**

定量研究基础:测量篇/(美)佩达泽(Pedhazur,E. J.),施梅尔金(Schmelkin P. L)著;夏传玲译. —重庆:重庆大学出版社,2013. 3
(万卷方法)
书名原文:Measurement, Design, and Analysis: an integrated approach
ISBN 978-7-5624-7231-5

Ⅰ. ①定… Ⅱ. ①彼…②夏… Ⅲ. ①定量分析 Ⅳ. ①O655

中国版本图书馆 CIP 数据核字(2013)第 026360 号

**定量研究基础:测量篇**

埃拉扎尔·J. 佩达泽(Elazar J. Pedhazur)
丽奥拉·佩达泽·施梅尔金(Liora Pedhazur Schmelkin) 著
夏传玲 译
张 彦 赵联飞 夏传玲 校
策划编辑:雷少波
责任编辑:李桂英 版式设计:雷少波
责任校对:谢 芳 责任印制:赵 晟
*
重庆大学出版社出版发行
出版人:邓晓益
社址:重庆市沙坪坝区大学城西路 21 号
邮编:401331
电话:(023)88617183 88617185(中小学)
传真:(023)88617186 88617166
网址:http://www. cqup. com. cn
邮箱:fxk@ cqup. com. cn (营销中心)
全国新华书店经销
重庆升光电力印务有限公司印刷
*
开本:940×1360 1/32 印张:7. 125 字数:217 千
2013 年 3 月第 1 版 2013 年 3 月第 1 次印刷
印数:1—4 000
ISBN 978-7-5624-7231-5 定价:25. 00 元

---

# 总序
# 社会研究方法的现状及其发展趋势

近年来,社会调查技术和社会研究方法都有很大的发展。在调查技术方面,自20世纪70年代以来,社会变迁多次横断面的跟踪调查研究,几乎成为所有国家和地区了解社会结构转变和社会发展状况的基础性调查。这种调查不仅对社会学的研究有很大促进作用,而且对整个社会科学的研究都产生了重大影响,并且这些调查结果有的已作为政府有关部门决策的重要依据。国际上比较著名的此类调查有:美国芝加哥大学全国民意调查中心(National Opinion Research Center,简称NORC)的“社会综合调查(General Social Survey ,简称GSS)”,英国埃塞克斯大学调查中心进行的“全国家庭生活和社会变迁调查”,法国经济和社会调查所进行的“全国经济社会调查”,日本社会学会组织进行的“全国社会分层与社会流动调查(简称SSM)”。中国台湾“中央”研究院社会学研究所,也每两年进行一次“台湾社会变迁基本调查”。美国的“社会基础调查”,现在已成为年度性的调查项目,它是美国国家基金会目前资助的最大的社会科学研究项目。以上这些调查,除美国的调查外,一般均因经费原因采用纵向的间隔性重复调查法,即每隔一段时间,进行一次全国规模的抽样调查。每次调查除保留社会研究所需的基本项目外,都有不同的主题。在间隔若干时间后,再重复同一主题的调查,这样的研究设计,使社会变迁研究在可以涉及更

为广泛的研究领域的同时,具有更好的积累性和可比性。多年来,这些基础性调查获得的资料,滋养着大批的社会科学研究者,有时一项调查就有很多名博士生用来写博士论文,以此取得的研究成就,其可靠性受到社会科学界的广泛认同。例如,1997 年出版,以台湾地区社会变迁基本调查数据为基础的研究报告集《90 年代的台湾社会,社会变迁基本调查研究系列二》收集论文 16 篇,内容涉及社会生活的各个方面,在台湾地区引起了极大的反响。

国内社会科学界在这方面也有了长足的发展。笔者所在的中国社会科学院社会学研究所的社会调查和方法研究室,组织或参与了多项与社会变迁有关的大规模抽样调查,取得了一定的研究成果,并积累了大量有关社会变迁的宝贵数据资料,其中主要有:

1. 城乡家庭变迁系列调查:该课题是由中国社会科学院社会学研究所牵头,联合北京大学和地方社科院的研究人员展开的一项类似多次横断面的城乡家庭变迁调查。这一调查始于 1981 年的“中国五城市婚姻家庭调查”,而后有 1988 年的“中国农村家庭调查”、1991 年的“中国七城市家庭调查”、1998 年的“中国城乡家庭变迁调查”。

2. 有关中国城乡社会变迁的系列调查:这一调查始于 1991 年的第二批国情调查,然后有 1992 年的“中国城乡居民生活调查”、1993 年的“第三批国情调查”、1995 年的“第四批国情调查”和 1997 年的“中国沿海发达地区社会变迁调查”。上述调查虽然还不是严格意义上的多次横断面的纵贯研究,但研究者已在研究设计中尽量考虑到纵贯研究的基本原则,如调查队伍的稳定、指标的可比性和样本空间的延续性等。

3. 中国城乡社会变迁调查:这一调查始于 2000 年,为中国社会科学院重大课题。目前已经完成第一期第一次调查和第二次调查,今后将把这一调查发展为连续的、定期进行的社会变迁调查。

在纵向调查技术取得长足进步的同时,20 世纪末至今,电话调查也有很大发展。电话调查涉及的范围几乎与个别(面对面)访谈同样全面。电话调查中使用的一系列方法,是在 20 世纪 70 年代后期和面对面调查一起发展起来的。在 20 世纪 80 年代中期,电话调查开始变得很普遍,并且成为许多场合中各种调查方法的首选。正如某些学者所言,一种在公共和私营部门被人们用来帮助提高决策效率的收集信息的有效方法为人们所普遍认同时,这一现象本身就具有方法论上的意义。不仅如此,电话调查还有很大的实

践意义,因为它为研究者提供了更多的控制调查质量的机会。这一机会包括抽样、被调查人的选择、问卷题项的提问、计算机辅助电话访谈(CATI)和数据录入。正因为如此,今天在各种社会调查中,如果没有发现其他重要的足以放弃使用电话调查的原因,电话调查由于其独特的对调查质量进行全面监控的优点,常常成为各种调查方式的首选。由笔者翻译,重庆大学出版社出版的《电话调查方法:抽样、选择和督导》一书,也于2005年面世。

无论是纵向调查抑或电话调查,实际上都是收集研究资料的方法,而应用社会科学的发展,不仅在于调查技术,即收集资料技术的发展,还在于研究方法和分析技术的发展。近年来,无论是定性研究方法,还是定量研究方法都有了长足的发展。

首先,计算机技术的发展可谓突飞猛进,它对当今社会生活的各个方面产生了巨大的影响,在悄悄地改变着社会科学的研究风格和研究方式的同时,也大大提升了社会科学学者的研究能力。这种影响表现在研究过程的各个阶段,从理论建构(概念映射)、问卷设计(专业的问卷设计软件)、调查实施(计算机辅助访谈、计算机辅助电话访问系统、网络在线调查系统)、数据录入(光学标记识别软件)到数据分析(包括文本、声音、图像资料的处理),甚至延伸到写作发表阶段。这样的过程发生在如社会学、经济学、政治学、心理学、教育学中,促进了学科之间的相互借鉴和交叉融合,至少在研究方法上呈现出这种趋势。随着计算机计算能力的大幅度提高,20世纪80年代后期,统计学领域内发生了一场“革命”,主要表现在对定类和定序变量的建模能力的大幅度提高上,以及与分布无关的统计分析模型的发展之上,特别是基于“Resampling”(包括Bootstrap、Jackknife、Monte Carlo模拟等)的建模技术。同时,计算能力的提高还带动了基于神经网络、动态模拟、人工智能、生态进化等新兴的分析和预测模型的发展。这些进展都为定量社会科学研究提供了更多的可供选择的工具。

亚德瑞安·E.拉夫特里(Adrian E. Raftery)依据社会学家所处理的数据类型,将定量社会学在美国的发展划分为三个时代:第一代起始于20世纪40年代,交互表是其主要处理对象,研究重点是关联度和对数线性模型;第二代起始于20世纪60年代,主要处理单层次的调查数据,Lisrel类型的因果模型和事件史分析是其研究重点;第三代起始于20世纪80年代后期,开始处理诸如文本、空间、社会网络等非传统的数据类型,目前尚没有形成成熟的形态。

拉夫特里的综述,虽然更强调定量社会学研究对统计学的贡献,但也大致勾勒出定量社会学在国外的发展脉络。

从分析模型的角度来看,定量分析在以下几个方向有了突破性发展:

1. 缺失值处理:由于社会生活的复杂性,社会调查数据常常出现缺失值,传统的处理方式是忽略这些缺失值,或者用均值替代。但现在则倾向于用多重插值法(multiple implation)或者其他基于模型的方法进行处理。这些技术的发展,不仅会增强我们对数据的处理能力,而且将改变我们设计问卷的方式。基于这些技术,我们在不增加被访者负担的前提下,大大增加了调查问卷的内容:每个被访者只回答问卷的一部分,然后通过对缺失值的处理,获得他们对未回答部分的估值。

2. 非线性关系:线性假定是经典定量分析的一个常见假定,但在实际研究当中,线性假定只能被看做是对社会现实的一个逼近和简化。面对具体的研究数据,如果没有理论上的明确指引(不幸的是,我们常常没有中程理论的指引),我们是无法在线性模型和非线性模型之间作出取舍的。但 MARS 模型的出现,让我们可以从经验数据当中获得最为拟合的变量之间的函数关系,而不必预先作出线性假定。这样,理论思考和数据分析就可以实现一个互动的循环过程,定量分析就不单单是对理论和假设的简单证伪过程,而是理论思维一个重要组成部分。

3. 测量层次:20 世纪六七十年代的统计模型,大多要求数据的测量层次在定距以上,如因素分析,但社会学的调查数据却大多为定类或定序数据。对应分析、Loglinear、Logit、Logistic Regression、潜类分析、Ordinal Regression、Normal Ogive Regression 等统计模型的出现,大大提高了定量社会学处理定类和定序数据的能力。

4. 测量模型:基于文化、社会、心理和认知等方面的考虑,在社会学界仍有人对问卷调查在中国的效度提出质疑。抛弃"本土化"的文化执著,我们更应当关注的是问卷调查的答题理论(item response theory),即被访者回答问卷题器时的过程模型。这方面的进展主要表现在两个方面:一是分解测量量表的成分,如 Rasch model、IRT 分析、Mokken 分析等;二是将测量模型与因果模型或其他分析模型结合在一起,明确把测量误差引入到分析当中,充分评估它们对分析结果的影响,如结构方程模型。

5. 潜变量模型:与测量模型相关联的另外一个发展方向是潜

变量模型，例如，潜类分析（latent class analysis）、潜结构分析（latent structure analysis）、潜预算分析（latent budget analysis）等。“潜变量”这一概念表明，我们可以通过测量“显变量”来测量无法直接观察的理论概念，如权力、声望、地位等。这样，理论和现实之间，通过“潜变量”到“显变量”的映射（测量过程），就有了连接的桥梁。

6. 分析单元的层序性：在定量分析当中，我们常常强调要避免出现“生态谬误”，即分析单元的层次和结论或推论的层次不一致。与其相关的方法论争论是“宏观和微观”的问题。随着多层次模型的出现，我们可以同时考察多个层次上的问题，我们可以把个人放在其家庭背景中，再把家庭放在社区的背景下，考察个人层次的变量对社区变量的效应，或者社区层次的变量对个体行为的具体影响。在定量分析模型当中，“宏观和微观”的连接获得了建模技术上的支持。在这个领域当中，还有一个方向也值得关注：分析宏观层次的数据，对微观层次进行推论。

7. 社会网络模型：区分“关系数据”和“属性数据”，是把分析重点从个体/群体等社会单元转移到这些社会单元之间关系的第一步，社会网络模型是目前发展较快的一个定量分析领域，其理论根基是结构主义。社会网络分析目前仍然具有较浓厚的“形态学”特征（基于图论的缘故），但却为我们理解社会关系在社会空间上的形态奠定了基础，通过计算机模拟和研究社会网络的历期数据，研究社会结构的“发生学”性质模型也处在萌芽状态当中。

8. 系统动力学：如果说社会网络模型是在社会空间上拓展定量社会学的研究手段，那么社会过程在时间上和物理空间上的属性，则是事件史模型、事件数模型、历期分析、Cox 回归、时间序列分析、Cohort 分析、状态空间模型等模型的研究对象。在这个领域，计量经济学为定量社会学研究提供了许多有益的范例。

9. 预测模型：上述模型仍然是在分析主义的范式下。有些社会学的应用研究，更强调模型的预测精度，而不是模型的认知价值，例如，社会趋势的预测。由于计算能力的提高，神经网络、基因算法、人工智能、模式识别等数据挖掘技术有了长足发展，已经出现了许多拟合经验数据的预测模型，比较成功的应用出现在计量经济学领域（如对股市的预测）。

10. 计算机模拟：对于社会学应用研究而言，研究的对象具有历史性、规模大、变迁的过程不仅漫长且表现某种渐进性的特点，且因社会隔离/社会伦理原因无法接近或有实验禁忌等，无法直接

进行观察和研究,这时计算机模拟就成为一个可供选择的替代方案。计算机模拟主要有两个类型:一是基于计算机网络的模拟:每台微机作为一个代理,整个网络作为"社会"实时演化,如法国的Swarm计划;二是基于概念模型的系统,在计算机时间上,按照既定规则运行,较有名的研究是罗马俱乐部的《增长的极限》,常见的软件有Simul,Arena等。自然科学家对此方向似乎比社会学家更有兴趣。

定性研究方法一直是社会学研究领域中比较传统的研究方法,在社会学研究的古典时期,它甚至是社会学家手中唯一的研究方法。但随着定量研究方法在社会学研究中的广泛应用,定性研究方法就似乎越来越不受人们的重视。但需要澄清的事实是,在定量分析模型取得飞速发展的同时,在过去的二十多年里,定性研究方法也有了长足的进步。主要表现在以下六个方面:

1. 研究素材日益扩大:除了传统的参与观察、深度访谈、专题小组访谈之外,会话、交谈、电视、广播、文档、日记、叙事、自传(autobiography)等社会过程中自然产生的素材,甚至社会学理论本身(理论的形式化),也开始进入定性分析的视野当中。所有这些资料,不仅可以以文本的格式存储,而且,新型的多媒体介质,如图像、声音和视频,作为原始的分析素材,也日益成为定性分析的新宠。

2. 分析方法更加多样:定性方法的种类在最近的二十多年中,更是有了一个质的飞跃。在比较传统的、源自语言学的方法,如内容分析、话语分析、修辞分析、语意分析、符号学、论据分析等方法之外,社会学家也创造出自己独特的定性分析方法,如施特劳斯(Strauss)等人的扎根理论、海斯(Heise)的事件结构分析、拉津(Ragin)的定性对比分析、Abbott和Hrycak采用最优匹配技术的序列分析、亚贝儿(Abell)的形式叙事分析(formal narrative analysis)、鲍尔(Bauer)等人的语库建设、Attride-Stirling等人的主题网络分析和神经网络技术应用的定性分析领域。所有这些方法的一个共同特征是,把定性研究向更加系统、更加精确、更加严格、更加形式化的方向推进。

3. 认识论基础更加多元化:现象学、释义学和本土方法论(ethnomethodology)的认识论,一直是定性分析的大本营,但近年来,实证主义也开始逐渐为定性分析所接纳,解释和阐释之间,由激烈的对立关系,逐渐演变为相互融洽的关系。

4. 研究过程更加客观规范:定性分析的一个主要问题在于阐释过程中不可避免的主观性。为了尽可能消除"解释者偏见"和主观选择性,定性分析开始遵循严格的程序模板或程序规则,并尝试引入定量分析中的"信度""效度""代表性"等概念,通过编码和对比,再加上传统的定性分析标准,如可解释性、透明性和一致性,使得定性研究的过程更加规范、阐释的结果更加客观,研究的结论更加可信。

5. 研究过程更加有效率:这主要应归功于大量计算机辅助定性数据分析(CAQDA)软件的涌现。从20世纪80年代以来,定性分析过程的数字化和计算机化,已经是一个不可逆转的大趋势。这种发展趋势与定性研究者的理论取向无关,不管他们的理论立场是实证主义、符号互动论,还是本土方法论,大多数定性研究者都在自己的研究当中,开始采用计算机来辅助定性资料的分析过程。据不完全统计,目前已经有二十多种定性分析的软件,分别隶属于德国、英国、法国、美国等国家。其中,有一些软件是国外研究机构的科研成果,可以免费使用,但比较成熟的定性辅助系统大多是商业软件。这些定性分析的辅助系统,不仅使得研究者从处理大量文字材料的繁复劳动中解放出来,而且能够让研究者共享他们各自分析的细节,从而改变定性研究的流程和研究集体之间的合作方式。同时,由于采用数据库结构,定性资料的管理也更加方便,这就为组织大型定性研究项目(包括多个研究地点、多个研究对象、历时的定性研究)提供了新的可能性。越来越多的定性研究人员开始走出他们的摇椅,坐到计算机屏幕前、湮没在访谈资料和故纸堆中的定性社会学家的形象已经一去不复返了。

6. 定性研究和定量研究的结合更加紧密:在定量分析方法的教材中,定性研究常常被看做是定量研究的前期准备工作,但定性研究者却持完全相反的观点,他们一般认为定性方法是自成一体的,可以完成从形成概念到检验假设的全部研究过程。在实际的应用研究中,定性方法和定量方法常常是交织在一起的,例如,克劳(Currall)等人在研究组织环境重要的群体过程时,通过内容分析把5年的参与观察资料量化,然后用统计分析来检验理论假定。格雷(Gray)和邓斯坦(Densten)在研究企业的控制能力时,利用潜变量模型把定性方法和定量方法有机结合在一起。雅各布斯(Jacobs)等人在研究比利时的家庭形态对配偶的家庭劳动分工影响时,首先用定量方法对纵向调查数据进行分析,从定量分析的结果中,又延伸出对核心概念的定性研究。这三个研究分别代表了定

量和定性方法相互融合的三个方向:①克劳等人的研究代表着定性方法的实践者试图将定性数据尽可能量化的取向,近年来涌现出的处理调查数据中开放题器的编码问题的工具软件(如 Words at,Smarttext 等,注意:它们都是由著名的统计软件公司出品的处理定性资料的软件),处理定性资料的传统内容分析软件(如 Nvivo、MaxQDA、Kwalitan 等)也开始提供将定性资料转换到常用统计软件的数据接口,这些工具上的革新将加快这种趋势的发展。②格雷和邓斯坦的工作代表了“方法论多元论”的取向,即在应用研究过程中,通过核心概念的测量模型,把定性研究和定量研究结合在一起。③雅各布斯等人的工作则代表了一部分定量研究者对过度形式化的定量方法的不满,并试图通过定性方法加以弥补。在定量研究领域中,对“模型设定”问题的关注,是定量方法重新试图返回定性研究这种取向的另外一种表现。

与社会调查技术和社会研究方法突飞猛进的现实相比,我国学术界在这些方面的论著的出版似乎显得有些迟缓。虽然已经翻译了美国的一小部分经典定量分析教材,如布莱洛克(Blalock)和巴比(Babie)的教材,也有自己编写的一些教材,如袁方等人的《社会研究原理和方法》、卢淑华的《社会统计学》等,此外,偏重软件操作的还有郭志刚的《社会统计分析方法——spss 软件应用》、郭志刚的《logistic 回归模型——方法与应用》、阮桂海的《spss for windows 高级应用教程》等。在《社会学研究》等专业杂志上,也常常有一些定量分析的应用研究,可是专门的方法和应用模型研究却没有,也没有专门的方法研究期刊。仅就定量研究方法的介绍而言,也存在一些缺陷,主要表现在:

1. 原理和操作脱节。
2. 过分依赖某些商业软件,不全面。
3. 与中国的实证研究相脱节。
4. 不能反映当前方法研究的最新进展。

与定量研究方法相比,由于各种原因,定性研究方法的引进和介绍都比较少。在福特基金会资助的方法高级研讨班上,曾讨论过一些定性研究方法。在定性方法研究方面也有少数专著,如袁方和王汉生 1997 年出版的教程,陈向明 2000 年出版的专著。但总体说来,我们对定性研究方法还停留在初步介绍的阶段,主要的介绍也局限在定性研究的研究设计和资料收集的阶段上,对定性分析方法的介绍,则没有能够反映出当代定性方法的最新进展。特别是在定性分

析工具（定性分析软件）的引进和研究上，基本上还是一个空白。虽然不乏一些出色的定性研究报告，但从方法研究上讲，我们才刚刚起步。当然，我们同时还应该注意到，在历史学领域，我国对定性资料的鉴别、考据和分析，积累了大量的经验和知识，这也应当是定性方法研究的知识来源之一，应努力发扬光大。

令人欣慰的是，社会研究方法的引进和出版方面相对滞后的状况终于有所改观。重庆大学出版社的编辑，以独到的学术眼光，逆当前出版界唯利是图的不良选题风气，投入了大量的人力、物力，组织出版"万卷方法"。自 2004 年至今，已引进社会科学研究方法方面的专著十余种，在我国社会科学界已经引起了一定的反响。然而，更为可贵的是，重庆大学出版社并未以已经取得的成绩而自满，而是再接再励，在原有"万卷方法"的基础上，进一步组织出版"万卷方法——社会科学研究方法经典译丛"。按我们的设想，"译丛"应该是一个开放的体系，旨在跟踪社会科学研究方法发展的前沿，引进和介绍这一方面的经典著作和最新成果。

"译丛"第一批有《抽样调查设计导论》《定量研究基础 · 测量篇》《定量研究基础 · 设计篇》《定量研究基础 · 分析篇》《问卷设计手册》《回归分析：因变量统计模型》《实用数据再分析》《社会网络分析法》《广义潜变量模型》《分类数据分析》和《复杂调查设计与分析方法》（书名也许有变化）等十余种，几乎囊括了研究设计、测量和分析方法的所有领域，涵盖从基础的回归分析到最前沿的潜变量分析和多水平模型等各种分析方法。无论是社会科学各专业的本科生、研究生，还是社会科学研究的学者都将从中有所收获。

"译丛"由中国社会科学院社会学所社会调查与方法研究室的多位研究人员担纲，主译者都是在社会研究方法各个领域中具有相当造诣的教师和研究人员。"译丛"的译者不仅仅把翻译看做是一个"翻译"，而且也把它看做是一次再学习和再创新。

我们期待"译丛"的出版能对社会研究方法的研究、应用和教学有所推动。

沈崇麟　夏传玲

中国社科院社会学所社会调查与方法研究室<br>社会调查与数据处理研究中心

# 译者前言

在众多的讨论社会科学方法的著作中，挑中这部专著介绍给中国的社会科学工作者和学生，主要是看重它关注了经验研究中的三个主要环节：测量、设计和分析。测量是通向经验细节之门，它是一切定量研究的基础和软肋，更是社会科学研究中最容易忽略的地方。对于社会公众而言，研究结论才是瞩目的焦点；而对于专业人士而言，测量才应当是需要窥视的法门。设计是把握研究全局的关键，它是俯瞰经验总体的高台，视野的大小、视角的转换、参照系的变动、经验和理论的通道、数据和命题的勾连，在这方面的功夫和修养，是方家和新手的试金石。分析是从经验重返理论的归途，在这里，我们需要把各种经验观察数据，在一定的参照系下，借助于各种统计分析模型，提炼成可以和各种理论命题对话的假设。简而言之，把握了这三个关键，各种方法、技巧和模型才会串联成一个总体：方法论。我们希望大家在阅读的过程中，仔细揣摩原作者谋篇布局中的方法论含义。

方法素养的高低，既是区分一个常人和一个专才的标准之一，也是我们鉴赏定量著作和论文的必备条件。但很多人在学习方法的过程中，常常采用的是最不得法的方法，因为他们把阅读理论著作的学习习惯带到了学习方法的过程中。对于理论学习而言，知识、理解、顿悟、联想和类比是关键；对于方法学习而言，理解只是第一步，把知识转变成自己的技能才是成功的标志。因此，方法不是“学”来的，而是“玩”来的，只有动手练习，才能把知识变成技能；没有动手练习，所有的方法都不过是纸上谈兵。在这一点上，我们希望大家都成为《射雕英雄传》中的郭靖，而不是黄蓉。前者是融各家功夫于一身的真英雄，而后者虽掌握各门各派的秘籍、识得各种招式，却只会耍嘴皮功夫，没有真本事。

将一部八百多页的“大部头”方法专著翻译成中文，其中的困难是译者最初所没有预料到的，不仅时间长达三年，而且翻译过程也由一个人的独奏变成一个团队的合奏，然后再变成一个人的煎熬。张彦老师负责了本书初译的组织和协调工作，并负责本书的一校任务。阚忠钰、刘杰、路小宁、欧珊珊同学分别承担了部分章节的初译工作。中国社会科学院社会学所的赵联飞博士承担本书的二校工作。在此，对诸位学术同仁的辛苦工作表示感谢。全书的三校和通稿工作由本人承担，虽竭尽心力，但限于知识领域和文字水平的限制，各种错误在所难免，还望诸位读者不吝赐教。

在这个漫长的过程中，我们得到了很多人的帮助，特别感谢中国社会科学院社会学所的沈崇麟研究员对翻译项目的指导和督促，感谢中国社会科学院社会学所的陈婴婴研究员、赵锋博士的不时指点，感谢重庆大学出版社雷少波同志在本书翻译过程中所给予的理解、谅解和鼓励。

夏传玲

2013 年 1 月 7 日

# 英文版前言

当我们开始写这篇序言时，我们不由得想起理查德（Ivor A. Richards）在他《文艺评论的原理》一书中的观察："分开来看，本书没有几条是原创的。在玩如此传统的游戏的时候，人们不应当期待出现一张新牌，真正重要的是出牌的手法。"（Richards，1926:1）在我们所涉及的主题中，要么在这一方面，要么在另一个方面，都已经存在很多文献。我们希望自己打出一副与众不同的牌，更重要的是，我们希望自己打出的这副牌能够对您有所帮助。

正如书名所表明的，我们尝试提出一个社会行为科学的综合研究方法。在第1章，我们将综述本书的内容、组织和取向，这里，我们将简要说明写作本书所要达到的目标。

总的来说，我们所涉及的主题，其他课本和课堂已经零零散散有所涉及。例如，在讨论统计学的教科书和课堂上，人们常常忽略理论问题和测量问题，在不知不觉之间，这造成了一种印象：仿佛它们不影响数据分析和结果阐释。在讨论测量的书籍和课程上，人们又很少或几乎不涉及设计和分析问题。在讨论研究设计的书籍和课程上，分析和测量问题常常是一带而过，或者完全不予讨论。

这种支离破碎的方法不可避免地带来一种结果：对研究探索的各个方面之间的相互关联和相互依赖，人们缺乏鉴识能力。它所造成的后果是，要求学生们熟悉"方法论"的各种期望，即使不被看做是一种虐待，也会被看做是毫无意义的事情（除了不得不写的博士论文或期末考试论文之外，有些学生并不想进入研究领域。对他们而言，尤其如此）。

我们并不是说，无论专业和理论兴趣何在，所有专业人员一定

要成为测量、设计和分析上的“专家”。但我们的确主张，这些研究领域以及它们之间的相互影响，对它们的一个基本理解，是成为研究结果的一个精明的消费者的必要条件；如果我们想要成为一个胜任工作的研究者，更是如此。因此，我们的目标是帮助大家精通研究的各个方面，而且，帮助大家培养一种视角，让大家能够关照它们之间的相互关联和相互依赖。同时，我们希望帮助您了解理论在研究工作中至关重要的指导作用。

尽管我们假定大家具有入门水平上的统计学知识背景（例如，了解“方差”“协方差”“简单方差分析”“相关系数”等概念），但我们还是会复习这些主题，然后再讨论更复杂一些的主题。如果大家细读本书的目录和第1章的话，大家就会发现，我们也讨论了一些高级主题。到目前为止，只有少数精通数学语言（这是表述这些主题的必要工具）的人们才能掌握它们。不过，随着计算机和软件的普及，即使是缺乏数学背景的人也有可能应用最复杂的分析方法。不幸的是，应用上的简便也大量增加了分析技术的误用和结果的错误阐释。为了帮助您毫不走样地学会我们所提出的分析方法，在讨论主题时，我们对计算机的输入和输出进行了大量的注释。

我们所涉及的这么多主题，我们怀疑能否在两个学期的课程中讲解完，更不用说掌握了。因此，我们希望这本书不是学期一结束就匆忙抛售的课本当中的一员。我们希望这本书成为大家的同伴，让大家时常回来，拓宽和深化大家对研究的理解。正是怀着这个目的，我们提供了大量的参考文献，我们相信，当大家在社会行为研究的广阔领域中追求知识时，这个参考书目将显现出它的价值。

最后，正如第1章所讨论的，本书的组织和表述方式具有很大的灵活性，无论是主题的选择，还是主题之间的次序，还是处理这些主题的复杂程度。我们希望这样的安排能够让教师按照自己的重点和学生的水平，因材施教。

# 致 谢

我们感谢纽约市卫生署流行病研究室的吉姆·吉鹏斯(Jim Gibbons),伊利诺伊大学的艾伦·科尼格伯格(Ellen Koenigsberg)博士和罗伯特·L. 林(Robert L. Linn)教授,纽约市教育委员会的伊丽莎白·塔勒普若斯(Elizabeth Taleporos),感谢他们对手稿的评论和改进所提出的建设性意见。

我们十分自豪、高兴地感谢海达尔·佩达泽(Hadar Pedhazur)(外号"马文电脑"),只要有必要降服我们的电脑、升级程序、按照我们的需求准备软件等,他就一定会来援救我们。

最深的谢意要献给拉里·厄尔堡姆(Larry Erlbaum),在结稿的日子不断延后的情形下,他所表现出的理解和耐心,他和作者的共鸣是图书出版中最好的传统。我们感谢阿特·利扎(Art Lizza)以他娴熟的技巧、敏锐和有求必应,指导了本书的写作过程。第二作者感谢霍夫斯特拉大学对她写作本书过程中所提供的支持。

埃拉扎尔·J. 佩达泽(Elazar J. Pedhazur)

丽奥拉·佩达泽·施梅尔金(Liora Pedhazur Schmelkin)

# 万卷方法总书目

万卷方法是我国第一套系统介绍社会科学研究方法的大型丛书,来自中国社科院、北京大学等研究机构和高校的两百余名学者参与了丛书的写作和翻译工作。至今已出版图书90余个品种,其中绝大多数是2008年以来出版的新书。

95 定量研究基础:测量篇
978-7-5624-7231-5

94 研究项目的实施:手把手指南
978-7-5624-6981-0

93 质性研究中的资料分析——计算机辅助技术应用指南
978-7-5624-6578-2

92 回归分析:因变量统计模型
978-7-5624-6976-6

91 倾向值分析:统计方法与应用
978-7-5624-6622-2

90 结构方程模型——SIMPLIS 的应用
978-7-5624-6603-1

89 在中国做田野调查
978-7-5624-6609-3

88 复杂性科学方法及应用
978-7-5624-6293-4

87 范式与沙堡:比较政治学中的理论构建与研究设计
978-7-5624-6375-7

86 心理学研究中的伦理冲突
978-7-5624-6131-9

85 社会科学方法论(国家十二五规划教材)
978-7-5624-6204-0

84 田野工作的艺术
978-7-5624-6257-6

83 图解 AMOS 在学术研究中的应用
978-7-5624-6223-1

82 应用 STATA 做统计分析(更新至 STATA10.0)
978-7-5624-4483-1

81 社会调查设计与数据分析——从立题到发表
978-7-5624-6074-9

80 质性研究导引
978-7-5624-6132-6

79 APA 格式——国际社会科学学术写作规范手册
978-7-5624-6105-0

78 如何做心理学实验
978-7-5624-6151-7

77 话语分析导论:理论与方法
978-7-5624-6075-6

76 学位论文全程指南
978-7-5624-6113-5

75 心理学研究方法导论
978-7-5624-5828-9

74 分类数据分析
978-7-5624-6133-3

73 结构方程模型:AMOS 的操作与应用(附光盘版)
978-7-5624-5720-6

72 AMOS 与研究方法(第 2 版)
978-7-5624-5569-1

71 爱上统计学(第 2 版)
978-7-5624-5891-3

70 社会科学定量研究的变量类型、方法选择与范例解析
978-7-5624-5714-5

69 案例研究:设计与方法(中译第 2 版)
978-7-5624-5732-9

68 问卷设计手册:市场研究、民意调查、社会

调查、健康调查指南
978-7-5624-5597-4

67 广义潜变量模型:多层次、纵贯性以及结构方程模型
978-7-5624-5393-2

66 调查问卷的设计与评估
978-7-5624-5153-2

65 心理学论文写作——基于 APA 格式的指导
978-7-5624-5354-3

64 心理学质性资料的分析
978-7-5624-5363-5

63 问卷统计分析实务:SPSS 操作与应用
978-7-5624-5088-7

62 如何做综述性研究
978-7-5624-5375-8

61 质性访谈方法
978-7-5624-5307-9

60 量表编制:理论与应用(校订新译本)
978-7-5624-5285-0

59 质性研究:反思与评论(第 2 卷)
978-7-5624-5143-3

58 实验设计原理:社会科学理论验证的一种路径
978-7-5624-5187-7

57 混合方法论:定性研究与定量研究的结合
978-7-5624-5110-5

56 社会统计学
978-7-5624-5253-9

55 校长办公室的那个人(质性研究个案阅读)
978-7-5624-4880-8

54 泰利的街角(质性研究个案阅读)
978-7-5624-4937-9

53 客厅即工厂(质性研究个案阅读)
978-7-5624-4886-0

52 标准化调查访问
978-7-5624-5062-7

51 解释互动论
978-7-5624-4936-2

50 如何撰写研究计划书
978-7-5624-5087-0

49 质性研究的理论视角:一种反身性的方法论
978-7-5624-4889-1

48 社会评估:过程、方法与技术
978-7-5624-4975-1

47 如何解读统计图表
978-7-5624-4906-5

46 公共管理定量分析:方法与技术(第 2 版)
978-7-5624-3640-9

45 量化研究与统计方法
978-7-5624-4821-1

44 心理学研究要义
978-7-5624-5098-6

43 调查研究方法(校订新译本)
978-7-5624-3289-0

42 分析社会情境:质性观察和分析方法
978-7-5624-4690-3

41 建构扎根理论:质性研究实践指南
978-7-5624-4747-4

40 参与观察法
978-7-5624-4616-3

39 文化研究:民族志方法与生活文化
978-7-5624-4698-9

38 质性研究方法:健康及相关专业研究指南
978-7-5624-4720-7

37 如何做质性研究
978-7-5624-4697-2

36 质性研究中的访谈:教育及社会科学研究者指南
978-7-5624-4679-8

35 案例研究方法的应用(中译第 2 版)
978-7-5624-3278-3

34 教育研究方法论探索
978-7-5624-4649-1

33 实用抽样方法
978-7-5624-4487-9

32 质性研究:反思与评论(第 1 卷)
978-7-5624-4462-6

31 社会科学研究的思维要素(第 8 版)
978-7-5624-4465-7

30 哲学史方法论十四讲
978-7-5624-4446-6

29 社会研究方法
978-7-5624-4456-5

28 质性资料的分析:方法与实践(第 2 版)
978-7-5624-4426-8

27 实用数据再分析法(第 2 版)
978-7-5624-4296-7

26 质性研究的伦理
978-7-5624-4304-9

25 叙事研究:阅读、倾听与理解
978-7-5624-4303-2

24 质化方法在教育研究中的应用(第 2 版)
978-7-5624-4349-0

23 复杂调查设计与分析的实用方法(第 2 版)
978-7-5624-4290-5

22 研究设计与写作指导:定性、定量与混合研究的路径
978-7-5624-3644-7

21 做自然主义研究:方法指南
978-7-5624-4259-2

20 多层次模型分析导论(第 2 版)
978-7-5624-4060-4

19 评估:方法与技术(第 7 版)
978-7-5624-3994-3

18 焦点团体:应用研究实践指南(第 3 版)
978-7-5624-3990-5

17 质的研究的设计:一种互动的取向(第 2 版)
978-7-5624-3971-4

16 组织诊断:方法、模型和过程(第 3 版)
978-7-5624-3055-1

15 民族志:步步深入(第 2 版)
978-7-5624-3996-7

14 分组比较的统计分析(第 2 版)
978-7-5624-3942-4

13 抽样调查设计导论(第 2 版)
978-7-5624-3943-1

12 定性研究(第 4 卷):解释、评估与描述(第 2 版)
978-7-5624-3948-6

11 定性研究(第 3 卷):经验资料收集与分析的方法(2 版)
978-7-5624-3944-8

10 定性研究(第 2 卷):策略与艺术(第 2 版)
978-7-5624-3286-9

9 定性研究(第 1 卷):方法论基础(第 2 版)
978-7-5624-3851-9

8 社会网络分析法(第 2 版)
978-7-5624-2147-4

7 公共政策内容分析方法:
978-7-5624-3850-2

6 复杂性科学的方法论研究(第 2 版)
978-7-5624-6396-2

5 社会科学研究:方法评论
978-7-5624-3689-8

4 论教育科学:基于文化哲学的批判与建构
978-7-5624-3641-6

3 科学决策方法:从社会科学研究到政策分析
7-5624-3669-0

2 电话调查方法:抽样、筛选与监控(第 2 版)
7-5624-3441-7

1 研究设计与社会测量导引(第 6 版)
978-7-5624-3295-1

# 万卷方法书友会

为了建设好“万卷方法”，更好地服务学界，重庆大学出版社组建了“万卷方法”书友会，凡购买我社万卷方法系列图书的读者，填写以下信息调查表或撰写万卷方法系列图书的书评，并通过 Email 发送到 wjffsyh@ foxmail. com 邮箱（重庆大学出版社　万卷方法书友会）即可成为书友会成员。我们将为各位书友提供以下服务：

●赠送人大经济论坛币 100 个。

●不定时发送有关学术活动（如研究方法培训班、研讨会）的信息。

●定期赠阅介绍新书动态、读书感受、方法学习、研究经验交流等主题的电子刊物。

●每本书前 50 名发来书评，且书评的原创内容（扣除引用原书及他人发言部分）不少于 400 字的读者，还将获得一本万卷方法的赠书。

●书评将选登于书友会电子刊物上，优秀书评还将推荐发表。

| 姓名： | 学校/单位： |
| --- | --- |
| 联系电话： | Email： |
| 论坛 ID： | |

## 人大经济论坛

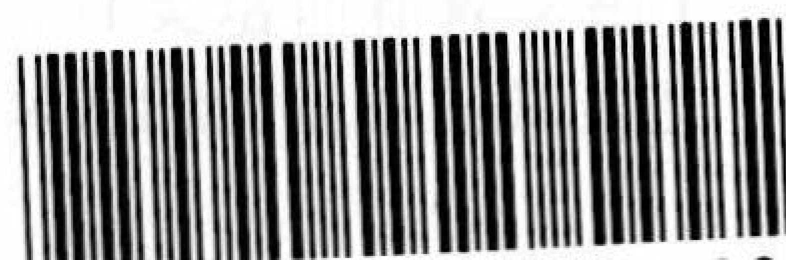
——国内最大的经济、管理、金融、统计类在线教育网站

人大经济论坛（网址：http://bbs. pinggu. org/）依托中国人民大学经济学院，于 2003 年成立，致力于推动经济学科的进步，传播优秀教育资源，目前已经发展成为国内最大的经济、管理、金融、统计类的在线教育和咨询网站，也是国内最活跃和最具影响力的经济类网站。

1. 拥有国内经济类教育网站最多的关注人数，注册用户以百万计，日均数十万经济相关人士访问本站。

2. 是国内最丰富的经管类教育资源共享数据库和发布平台。

3. 论坛给所有会员提供学术交流与讨论的平台，同时也有网络社交 SNS 的空间，经管百科提供了丰富专业的经管类在线词典，数据定制和数据处理分析服务是您做实证研究的好帮手，免费的经济金融数据库使您不再为数据发愁，更有完善的经[illegible]，只要您是学习、研究或从事经管类行业，人大经济[illegible]



# 目　录

# 1 概　述

在社会行为科学中，许多学生和专业人员都缺乏必要的背景知识，在所感兴趣或所在的专业领域中，他们因此而无法让自己变成研究文献的聪明消费者，这是一件令人遗憾的事情。而且，在一个专业化的时代，专业人员（学生也一样）从事研究的时候常常求助于"方法论顾问"的"服务"。在大多数情形下，这些专家的意见是必要而且有益的，但是，这些意见常常和专家处方混淆在一起。许多研究人员和博士生相信，分析数据、阐释结果、从中引申含义等"任务"，可以委托给这些"顾问"来做，而且，这也没有什么不合适的地方。但盲目服从顾问的处方或者委托他们进行数据分析和结果阐释，无异于让他们代替自己思考。

这种做法带来的有害结果之一是，出现了一种完全放弃自己的批判能力的趋势。许多专业人员相信，他们的角色不是评估研究报告，而是了解其发现、结论和含义。这样，接受或拒绝一个研究发现及其含义，就不是基于周详的判断，而是基于其他方面（例如，常识、研究者的地位）。对社会行为科学广阔领域中的实践和政策决策而言，我们希望大家能考虑这种状况所带来的潜在有害效应。

## 思至上

一个众人皆知但值得重复，甚至要求重复的事情是：若想有意义，任何活动（包括阅读研究报告）必须先从稳健的、批判的思考开始。如果有一个需要我们在一开始就发送给大家的短信，那么，它一定是：运用大家的常识，不要让胡言乱语和技术俚语战胜大家。

对方法和量化的过度依赖会带来一个副效应,即研究人员和研究消费者都在丧失批判性思维。特别是经过计算机的磨光处理之后,专业术语、公式和精致的分析都会散发出一种诱惑,一种几乎魔幻般的品质,它们很有可能让人们对它们的真实含义丧失注意、丧失思考。

毋庸置疑,如果我们想批判地评估一份研究报告,我们就必须掌握所使用的各种方法和分析思路。不过,为了强调应用常识的重要性,在开篇的这个阶段,我们想举出一些示例,其中,运用常识就足以怀疑作者的论述。试想一下,朗(Vonda O. Long)说明她在研究中使用一个测量工具的下列“理由”:

> “尽管 BSRI(贝姆性别角色量表)已经受到批评……但它仍然得到广泛的应用。”(Long, 1983: 324。在另一篇论文中,她使用了同样的“理由”,参见 Long, 1989: 85)。再试想一下,芬哈默(Adrian Furnham)说明她在研究中如何挑选量表:“挑选的依据是它们的稳健性和心理测量学上的满意度。”(Furnham, 1984: 283)就其中的一个量表,芬哈默写道:“我们认为,它已经是一个稳定、有效且经济的量表,且应用于许多研究当中。”(Furnham, 1984: 284)她把其他量表刻画为具有“令人满意的心理测量的结构”(Furnham, 1984: 284)和(或)“得到广泛应用”(Furnham, 1984: 284)。

上述引文就是我们所能了解的,她所采用的各种度量的全部属性。我们希望大家能看清楚这类陈述的空洞性,即便大家对测量所知甚少。

下面是另一个不同类型的例子。在讨论“图式参照对社会行为的后果”这份实验报告中,桑德兰兹(Lloyd E. Sandelands)和考尔德(Bobby J. Calder)写道:

> 我们首先检验的是,在自参照条件下,这些词汇更难出现或更不常见的可能性。采用桑代克和洛尔格的词频常模(Thorndike & Lorge,1941),所有呈现给被试的词汇,依据它们在日常语言中的出现频率,我们都进行了编码。(Sandelands & Calder, 1984: 761)

一个 20 世纪 30 年代末建立的词频常模对 20 世纪 80 年代实施的一项研究的恰当性,一点思考就足以让我们产生怀疑,如果不

是立即拒绝的话。显然,这两位作者也感觉到,无论多么不相干,他们也需要一个准则,以免人们批评他们提出了没有理论基础的命题。实际上,服从科学协议的规范是如此根深蒂固,以至于只要引用一篇论文,无论多么不相干,似乎给这个论述涂上了科学的严谨和客观的脂粉,即使在审稿人和编辑的眼中,也是如此。

重申一次:为了成为一个研究的聪明消费者,更不用说为了成为一个胜任的研究人员,在研究探索的各个方面,积累知识和技能都是必不可少的。但是,如果大家不思考,再娴熟的技术也于事无补。

下面我们简要介绍一下本书的内容、组织和取向。

## 内 容

测量、设计和分析是本书所讨论的主要领域,讨论它们的书籍和文章可谓汗牛充栋。因此,不言而喻,我们的论述绝不可能巨细无遗。在这一节中,我们将概述本书所选择的主题,对选择的理由也会作一些概括性的评述。

### 测 量

测量是社会行为研究的阿喀琉斯脚踵。在社会行为科学中的大多数项目(特别是博士培养项目)中,尽管我们要求学生学习少量的统计学和研究设计,但这种要求极少能达到测量所关注的程度。这样,许多学生得到一个印象,开发和使用测量并不需要特殊的技能。因此,在很多研究中,他们几乎不关注测量的属性,这也就不足为奇了。令人遗憾的是,许多读者和研究人员没有能够认识到,无论理论表述如何精深,设计如何复杂,分析技术如何精致,都不能弥补粗劣的度量。

有许多书籍和众多文章讨论测量的各个领域,但它们讨论的范围、深度和复杂性各不相同,有些是一般的导论性质的综述,有些讨论的是宽窄不等的特殊主题。例如,有许多书籍和文章讨论成就、智力、态度和人格的测量,这些仅仅是沧海一粟。也有一些书籍和论文讨论测量理论、测量模型、心理测量理论或诸如此类的特征。某些主题可能是特定领域中所独有的,其他主题则或多或少需要推敲,依赖具体的语境。例如,诸如多选题与论文考试、猜题、评分实践、参照准则与参照常模的测试、测试的等价性等主题,

主要（如果不是唯一）出现在致力于成绩测试的讨论中。投射技术、应答方式、应答集等，很可能是在讨论人格、态度等测量的文章中得到广泛讨论。测量理论也或多或少有些差异，差异的大小取决于所研究的特定主题领域（例如，成绩、能力、态度、人格）。

上述评论应当足以说明我们的主题选择、范围覆盖、陈述层次等方面的理由。我们选择主题的主要依据是它们在研究探索中的角色以及它们的共性或普遍性。

第 2 章专注于测量在科学研究中角色的一般性介绍，包含的主题有测量的定义、测量量表，以及测量和统计学之间的关系。

效度是社会行为测量最重要的主题之一，因此，我们用两章来讨论它：第 3 章集中讨论准则相关的效度验证，第 4 章集中于建构效度验证。

第 3 章讨论的主题包括准则的定义、准则的性质和类型、预测、预测效率和区别预测。第 4 章以探讨建构的含义以及建构和指标之间关系开篇。然后，我们在三个标题下探讨建构效度验证的各种流派：①逻辑分析，其中，我们讨论了建构定义、题器内容、测量和计分程序；②内结构分析，其中，我们对探索性因子分析和证实性因子分析进行了直观的介绍；③跨结构分析，其中，我们讨论了趋同验证和判别验证的概念，如何用多特质、多方法的矩阵方法来评估它们。本章以我们对内容效度的评注结尾。

第 5 章讨论估计信度中的理论和实践考量。包括的主题有信度在测量和研究中的地位、信度概念、经典测试理论及其变更和扩展，强调内在一致的信度估计方法、效度和信度之间的关系、低信度所带来的逆效应。

第 6 章是对社会行为研究一些选定测量方法的导论，分为下列主题：①评分量表；②语义微分；③访谈；④观察。和它们当中的每一个有关的问题，包括建构、应用、分析、阐释和偏差来源。

## 设　计

第二篇的开篇是对科学和科学研究的一般介绍（第 7 章）。涉及的主题包括基础研究和应用研究，自然科学和社会行为科学之间的异同，社会行为研究结论和政策建议，社会行为科学的内容和方法等。

第 8 章讨论定义和变量。就定义而言，我们讨论了定义在科学

研究中的作用、定义优劣的标准、理论定义（一般而言，特别是社会行为研究）和经验定义。然后，我们讨论了变量的定义，并从测量和设计的角度探讨了变量。

理论、问题和假设这三个相关的话题是第9章讨论的主题。在对有关理论的定义进行一些评述之后，我们就开始关注理论在科学研究中的主要角色。讨论的问题包括理论和事实、理论作为参照系、理论的偏见效应。然后，我们又讨论了科学研究中的证实和证伪，以及科学的进步。最后，我们考察了社会行为科学中的理论状况，本节以对社会行为研究的短暂性的观察结束。

讨论问题的章节以“构成科学研究中的问题是什么”的讨论开始，续之以问题表述的不同形式。问题形式和理论表述之间的关系，以及它们对设计类型和所应用的分析的含义，我们也进行了讨论，并举例予以说明。然后，我们讨论了问题的理论意义这一复杂问题，讨论了可研究和不可研究的问题，讨论了以前的研究在问题表述中的角色等。

讨论假设的章节以讲解和讨论假设开始，假设的形式和前面所讨论的问题表述形式是平行的，续之以讨论假设的指导和误导（即无法证实）作用，以及检验从不同理论观点所推导出的备择假设的作用。我们还另辟一节专门讨论假设的统计检验，讨论的主题有围绕统计检验的争论，P 值的阐释和误释，统计显著性和理论重要性之间的区分。本章的结尾将讨论基于决策的假设统计检验方法。

第10章专门讨论研究设计的基本原理和观念，讨论了两个相关的主题：控制和效度。继讨论控制在科学研究中的关键角色之后，我们描述了各种控制形式，以及它们和所考察的研究设计类型之间的关系。

然后我们再讨论更宽泛的主题，即效度。在简要的概述之后，本章的剩余章节就致力于讨论内在效度和外在效度。按照复杂程度的不同，我们将有详有略地讨论威胁内在效度和外在效度的不同因素。

第11章是第10章的补充，集中讨论社会行为研究中的常见的人为假象和陷阱，以及它们对研究结论效度的威胁。本章的组织围绕着两个主要的人为假象和陷阱来源，即被试和研究者。和第10章一样，主题讨论的深浅程度不同，这取决于主题的普遍性和

(或)复杂性。

接下来的三章分别讨论不同的设计:第12章讨论实验设计,第13章讨论准实验设计,第14章讨论非实验设计。总的来说,这些章节包括所考察的设计要素的定义和细节,它的特质、优势、弱势,特别是对效度的含义,以及研究示例。

然后,我们将讨论一些初级设计,并在入门的级别上建议一些分析方法。对于所推荐的每一种方法,我们给出了本书第三篇的相应章节作为参考,在这些章节中,我们将详细讨论这些方法(参见下面的“谋篇”一节)。

第二篇的最后一章是对抽样的导论(第15章)。这一章的开篇讨论样本和抽样策略的定义,然后再讨论抽样的目的和优势。在区分非概率抽样和概率抽样之后,我们的讨论限于后者,讨论的主题包括估值的属性、抽样分布、所选择的抽样策略、效应规模,以及统计力分析和决定样本量大小之间的关系。

## 分　析

在讨论第三篇内容之前,我们将首先就分析作一些粗浅的说明;然后,表明我们对读者在此领域的背景假定;最后,我们将就本书对分析技术的讨论范围和讲述方法进行一些说明。

### 开场白

高尔顿(Francis Galton)在“统计学的魅力”一章中曾经说过:

> 有些人憎恨“统计学”这个名字,我却发现它充满美丽和趣味。只要我们不粗暴对待它们,而是施以更高级的方法谨慎处理、小心阐释,它们就会展现其处理纷繁现象的非凡能力。(Galton, 1889: 62)

正如我们在前言所提到的,统计学的讲解常常不注意理论背景、应用的设计特征或所用度量的属性。如此肢解之后,出现不当阐释、反感和粗暴对待统计学的现象,也就不足为奇了。

前面我们已经提到盲目听从专家在数据分析和结果阐释方面的建议、没有应用常识的一些逆效应。如果提供建议的“专家”不熟悉或者不理解理论问题及其过程,问题将会变得更严重。有些专家的所作所为仿佛是说,理论问题、设计问题和测量问题互不相

干，人们唯一所要了解的就是，哪些是 X，哪些是 Y。面对这种潜在的误用，作为现代"研究设计和分析"概念的主要奠基人之一，费舍尔爵士（Ronald Fisher）曾经这样说：

> 统计学家不能推卸一种责任：他应当理解应用或建议领域中的过程。我的观点就是，我们可以分离（应用领域中）所牵涉的问题和统计学家技艺中、严格意义上的技术问题，一旦作出这样的分离之后，各种问题就只是正确应用人类的推理力的问题，这是所有智者（希望自己聪明的人）都同样关切的，而统计学家（作为统计学家）对此并无特殊的权威。在科学推理的原则上，统计学家不能免除保持清醒头脑的责任，正如其他思考的常人同样不能免责一样。（Fisher，1966：1-2）

通过上述引文，我们希望大家明白一个道理，选择一种分析方法绝不是一种机械的程序。在选择分析方法时，应用研究的所有方面（例如，理论表述、设计特征、测量工具的属性）都应纳入到考虑的范围。因此，当专家们对"正确"分析的意见不一致的时候，特别是在相对复杂的研究中，我们不应该感到丝毫的惊奇。米拉夫斯基（J. Ronald Milavksy）等人选择"正确"的分析方法的艰辛过程，就是一个很好的示例。在一个研究电视和攻击性之间关系的大型研究中，为了选择一个分析方法，在大约三年的时间内，他们就"最佳"的分析方法咨询了一些学者。被咨询的学者在有一点上是一致的，即米拉夫斯基等人的分析方法是不恰当的，但在什么是恰当的分析方法上，他们之间没有达成一致，而且是很大的不一致。米拉夫斯基等人说，事情总算得到解决，当：

> 我们最后认定结构方程模型是最佳方法。然后，就是购买 LISREL 软件，让它在计算机上运行，学习如何使用和阐释结果。当数据分析真正开始时，我们还是重新做了所有前期的交互表分析。[1]（Milavsky et al，1984：183）

一个研究的不同方面和所采用的分析方法之间的相互关系，我们将分别在不同的章节中加以讨论，同时，我们也将说明，对相同的数据应用不同的分析方法时，我们将得到不同的结论。在这里，我们只请大家关注邓肯（Otis Dudley Duncan）关于不同分析方

---

① 围绕 LISREL 软件的结构方程模型的介绍，参见第 23 和 24 四章。

法的效应一个注释。他和学生们用更“严格”(恰当?)的技术,重新分析了已经发表的研究数据,结果发现:

> 几乎不变的事实是:①当使用严格的检验时,原作者所主张的关系并没有得到数据的恰当支持;或者②原作者忽略了相同数据中的重要关系;或者更可能是③两者兼而有之。(Duncan, 1978:404)

### 对读者背景的假定

我们假定读者有初级统计学知识,即对基本概念和方法(例如,平方和、标准差、标准误、t 和 F 比值、简单方差分析、相关分析)有一些了解。当然,我们急于加上一句,我们并不严格坚持这个假定。讲授研究设计课程的结果显示,两个学期的统计学课程是一个必需的条件,我们相信,那些没有学过“传统”统计学课程的读者,将获益更大。或许正是因为这类课程是在理论和设计的真空中教学的,学生们根本就没有抓住“要点”。而且,我们有一个强烈印象,很多学生的反应像听天书似的,仿佛他们一点也不熟悉讲授的主题。姑且不论其技术含义,我们的另一个印象是,当我们用均值、方差、变量或相关系数来讨论分数的变异度时,许多学生的反应是,这些术语在英语中毫无意义。

有鉴于此,我们认为有必要重温一下基本的统计学概念。如果大家恰恰需要这样的复习机会,我们希望大家给统计学第二次机会,以不同的眼光重新审视统计学,我们甚至奢望大家能喜欢统计学! 如果大家没有这样的需求,我们希望大家能理解我们复习的动机,跳过下面大家已经完全熟悉的章节就是了。

### 讨论的范围和方式

我们所讨论的概念和分析方法的范围比较广,涵盖最基本(例如,方差、协方差)到最高级(例如,结构方程模型)。有的人可能会批评这种一袋装的方式,有的人则可能会质疑说,有一些方法还是被忽略了(例如,列联表分析、时间序列分析)。下面,我们就方法的选择和方法的讲述方式作一个说明。

我们把注意力集中在第一篇和第二篇所讨论的主要论题和主要设计相对应的分析方法上。例如,我们将从概念验证(第 4 章)

的角度解释因子分析，并给出示例（第 22—23 章）[1]。同理，和第 12—14 章所讨论的各种设计一致，我们讨论了连续自变量和定类自变量回归分析的各种不同应用（第 17—21 章）。

贯穿始终的一个关注点是应用，即分析和其他研究方面之间的契合。我们认为，实例是应用类教学的最好方法。在实例中，方法变得生动，特别是对那些没有数学倾向，且主要（假如不是唯一的话）的关注点是应用和阐释的人。在应用一种方法的过程中，人们才能更好地理解这种方法和既定问题、情境的相关性，并因此学会如何阐释结果。实际上，“和性爱一样，学习方法，示范总会强于讨论”（Leamer，1983：40）。

因此，所有主题和方法都是在一个简单数据实例的背景下讲述的，我们会详细分析实例，详细阐释结果。下面，我们就数据实例作几点说明。

尽管我们希望把数据实例放在一些实体性的研究背景下，但形式必然是简略的，也就是说，给变量添一点味道、给假设的模型一个粗略描述。显然，像这样的一本教材，不可能进入详细的理论考量。如果有必要的话，我们会提供研究主题的一些参考文献。无论何种情形，我们都不断地提醒大家，我们所采用的具体实例，仅仅是演示，我们从没有主张或暗含它们的理据或效度。

不可避免的是，实例的选择受到我们研究兴趣的影响，我们或多或少熟悉的实体性研究领域包括一般心理学研究（特别是社会心理学）和教育研究（特别是关于最广泛意义上的学校效应的大型研究）。如果大家不中意我们的实例，就请大家用自己感兴趣的领域中的实例替换吧。

我们采用较小，因此也不太符合现实的数据实例，以避免计算技巧的牵制。即使是采用计算机时，我们也采用这类实例。这样，我们就能够通过手工计算把它们和当前主题讨论中的公式和方法联系起来，来演示如何得到结果。我们相信这种教学方法有助于运用具体且易处理的实例，来阐释一个计算机程序输出结果的各种特征。

被动跟随一种方法或技术的讲座，常常会把人诱导到一种信

---

① 本书第 7—15 章见《定量研究基础：设计篇》，第 16—24 章见《定量研究基础：分析篇》。全书后同。

念中,自认为自己已经了解这种方法,但一旦要用的话,自己还是不知道该如何操作。如果一个人试图操作一种自认为熟悉的分析方法,却发现自己从哪里下手都还不知道,或许没有什么事情比这更令人清醒的了。为了避免出现这种窘境,我们强烈推荐大家重复我们的分析,再做一些自己的实例。正是这个原因,我们也包括了学习建议,给大家实习已经学到的既定分析方法的机会。我们建议,在学习的初始阶段,大家既用手工计算,也用计算机技术。当然,一旦大家比较熟悉一个既定分析技术背后的思路后,大家就可以把计算的苦差事交给计算机了。

## 内　容

在各种分析方法的讨论中,计算机分析是一个主角。因此,在第3篇的第1章(第16章),我们着力于介绍计算机及其程序。这一章的开篇是关于计算机的一些基本观察,我们讨论了计算机数据分析的优势,有关计算机的一些错误概念,以及迷信计算机和程序绝对不会出错所带来的负面效应。随后,我们讨论了统计软件的基本构成,对如何评估统计软件给出了一些建议。紧跟着,我们列出了本书所用程序的选择标准,以及所用统计程序包的目录。然后,我们讨论了使用统计软件时所应注意的一些问题。随后涉及的主题有手册的使用,程序默认值、输入文件的生成和编辑、出错提示。在下一节,我们简述了本书对输入和输出文件所采纳的惯例,同时也说明了我们对输入和输出文件进行注解的性质和目的。在最后一节,我们介绍了本书中所用的每一个统计程序包,也描述了使用它们时所采纳的惯例。

在第17章,我们将讨论简单回归分析。一开始,我们回顾了基本的统计概念(例如,方差、协方差)。然后,我们讨论了简单回归分析的元素(例如,回归方程、平方和的分解)。随后,我们讨论了在回归分析中作图的优势,也列举了计算机程序所生成的图形。在下一节中,我们着重讨论回归分析的显著性统计检验。然后,我们又详细讨论了回归模型背后的假定,回顾了违背这些假定所带来的效应。在下一节中,我们着重讨论模型诊断,这一节又分成残差分析和影响力分析两个小节。在这两个小节中,我们用数据示例讨论和演示了某些主要的方法。本章用我们对相关系数模型的注解结尾。

探索性研究中的多重回归分析是第 18 章的主题。[①] 实际上，在两个自变量的情形下，我们很容易理解输出结果的分析和阐释等方面，而且，在这种特殊情形下，手算也比较简单，因此，我们用了相当大的篇幅来讨论手算。然后，我们讨论了两个自变量以上的情形，给出了计算机分析的示例，以及我们的注解。

鉴于对方差分解的错误理解和错误应用很普遍，特别是在多层次回归分析的形式中，我们花了整整一节的篇幅来讨论方差分解。然后，我们讨论了效应衰减以及它和样本量之间的关系。紧接着，我们讨论了多重共线性和它的逆效应，并给出了一些补救措施。本章以讨论曲线回归分析的一节结尾。

第 19 章着力于讨论如何分析含一个定类自变量（例如，不同的处理、婚姻状况、种族）的设计。我们首先讨论了包含在定类变量中的编码信息的概念，然后，说明了如何在回归分析中使用这些编码信息。我们列举了三种编码方案，讨论了每一种编码方案的独特特征，并用相同的数据示例，演示了它们的应用。

然后我们讨论了均值之间多重比较的主题，并用事前对比和事后对比作了演示。我们介绍了如何使用相关的编码方案，直接从回归分析中获得这类比较的方法。

在本章的最后一节，我们讨论了在含一个定类自变量的设计中，回归分析和简单的方差分析之间的相似性。以此来说明，尽管存在分析术语和技巧上的差异，这两种分析方法实际上殊途同归。

在第 20 章，我们把第 19 章中讨论的概念和方法，扩展到多个定类自变量或因子设计的情形，并说明如何把第 19 章所讨论的编码方案应用到多个定类自变量的情形。我们用含两个定类自变量或两因素的设计，来回顾并演示因子设计的优势，特别是这种设计在侦测自变量对因变量效应的互作效应上的优势。除了这些方面之外，我们还说明了“集中比较”的概念，用来寻找一个互作效应。我们演示了如何利用回归分析的输出结果，用手工进行计算，或者通过相关的子命令用计算机进行计算。

然后我们讨论了在因子设计的背景下如何表述假设，并用预期出现互作效应的情形和预期没有互作效应的情形作了说明。之

---

① 我们将在不同章节讨论解释性研究和预测性研究之间的区分（即第 3 章、第 8 章、第 10 章）。

后,我们回顾了如何分析两个以上因素的设计。我们用三因素的一个设计,作为研究示例,来阐述一些基本思路。

之后我们将转而讨论非正交设计(即组频次不相等的设计)。在讨论组频次不相等的实验设计和非实验设计之间的关键差异的同时,我们还分别讨论了这两者各自有关分析和阐释的问题。本章最后一节专门讨论多重回归分析和方差分析之间的异同。

在第21章中,我们讨论如何分析含定类自变量和连续自变量的设计。因此,我们把第17—20章中分别讨论的方法,融合在这一章中。我们讨论了两种基本设计,并用示例讲解了对它们的分析,即属性—处理—互作效应和协方差分析。我们将说明,这两种设计之间的差别,不在于分析的方法,而在于研究的关注点。当我们的研究关注点是探讨个体属性和实验处理之间在对因变量的效应上是否具有互作效应时,我们就会把这种设计看做是属性—处理—互作效应。相比之下,如果研究的关注点是控制个体属性、探讨实验处理的效应时,我们就把这种设计看做是协方差分析。但是,在这两个示例中,这些分析只适用于实验设计所获得的数据(参见第12章)。

然后我们讨论了如何在准实验设计(第13章)中应用协方差分析,来校正非等价组之间的差异。除了讨论这种应用所引发的严重的设计和逻辑问题之外,我们还讨论和演示了协变量测量误差所带来的偏差效应。然后,我们回顾了准实验设计中替代协方差分析的其他方法。然后,我们简述了把这些分析方法推广和扩展到其他设计(例如,间断回归、因子设计)的问题。

探索性因子分析是第22章的主题。我们一开始就讨论了因子分析以及其和理论的关系,然后,我们对因子分析中所用的矩阵属性作了一些注解。我们用了两个数据示例,一个示例中,因子是相关的;另一个示例中,因子是不相关的。我们分析了这两个示例,并作了注解。讨论的中心自始至终都是如何在建构验证的过程中应用因子分析。我们讨论的主题包括主成分分析和因子分析之间的区别,公因子方差和独特性,正交转置和斜交转置,因子矩阵、再生的相关系数,因子的阐释和命名,抽样和样本量、因子分、基于因子的量表构建,因子分析结果的报告等。

在第23章中,我们讨论证实性因子分析(CFA)。简短的讨论之后,我们把证实性因子分析看做是结构方程模型的一个子模型,

而结构方程模型是第 24 章的主题。然后,我们介绍了一下 LISREL,这是我们将在第 23 章和第 24 章将要用到的计算机程序。在介绍中,我们涉及的主要话题包括在 LISREL 子模型中的矩阵,用来进行证实性因子分析,LISREL 的控制语句、默认值和模型设定。用于探索性因子分析的相同的数据示例(第 22 章),也同样用于证实性因子分析中。在对输出结果的注解中,我们涉及的主题包括拟合指数和模型检验有关的逻辑、方法和假定。

在下一节,我们介绍了 EQS,这是用于第 23 章和第 24 章中的另一个计算机程序。我们回顾了 EQS 所用的术语、输入和控制语句。以前用 LISREL 分析的数据示例之一,再用 EQS 进行了分析。在对输出结果的注解中,我们讨论了 EQS 的特色,以及它和 LISREL 之间的异同。然后,我们讨论了证实性因子分析背景下,如何分析多特质多方法的矩阵。我们分析了一个由三特质三方法构成的矩阵示例。在分析的过程中,我们讨论的主题包括:嵌套模型的检验,模型的修正,为了修正模型而使用程序输出结果中的一些指标,把方差分解为源自特质、方法和误差的各种成分。

在本章的结尾,我们讨论了第 5 章中所讨论的测量模型,我们分别考察了平行模型、真值等价模型、同属模型。我们用一个小数据示例来演示这些模型的检验。在第 24 章,我们讨论了结构模型的表达、估计和阐释。结合概念、定义和理论的关系、设计类型等,我们讨论了因果律这一广受争议的主题。

然后,我们介绍了如何用 LISREL 来分析结构方程模型。介绍的方式和第 23 章一样,我们说明了结构模型所用的矩阵,并作了演示。然后,我们说明了输入和控制语句。同样,我们也介绍了 EQS,也演示了其应用。

我们分析了几个数据示例,既用了演示数据的第一部分,也用了整个数据集。使用数据第一部分的目的,是为了讨论和演示不同模型的分析和阐释。结合对各种模型的分析,我们也讨论了一些主题,包括使用单个指标与多个指标,直接效应和间接效应,标准化解和非标准化解。我们用对所选择主题的简要注解来结束这一章。

## 组 织

退一步说,有关本书章节组织的决策是一项挑战。我们寻找

一种结构,它既能承载如此宽泛的主题的有序表述,又不能忽略它们之间的相互联系和相互依赖,在这个过程中,我们考虑了几种不同的方案。尽管我们知道,大家可能认为我们所选定的结构不是最佳的,但我们希望大家会发现这种结构是有用的。为了增加后一种可能性,我们将简述本书组织的主要方面,也给大家如何使用本书提供一些建议。不过,归根结底,将本书材料按照最适合大家的需求和目标的方式来加以组织的人,正是大家自己。

我们相信,从上一个章节关于本书内容的描述中,本书的总体结构已经显露无遗。在下面的章节中,我们将就本书组织的特定方面作一些说明。

## 重复主题

在不同的水平上,从不同的视角来考察重复的主题,这可能是刻画本书结构的最佳方式之一。大多数情形下,我们首先在一个直觉的层次上描述每一个主题,然后再用更形式化、更严谨的方式加以讨论。在很多情形下,我们将从本书的主要视角(即测量、设计和分析)中的一个以上的视角来考察相同的主题。

我们相信,一个示例将说明我们的意图。我们首先在直觉层次上,从测量的视角来引入"指标"概念,聚焦于建构效度(第4章),然后,我们又从设计视角讨论了指标(第12章和第13章)。在上述章节中,有关分析的问题仅在一般的意义上有所涉猎,但参考文献留给了第三篇(即第23章和第24章),在那里,我们讲述了有关的分析流派。

由此可见,作为保持整体性同时简化表述的方法,我们诉诸频繁的交互参考,并建议大家在学习一个既定主题时,精读本书的各个章节。例如,讲述相关分析流派的参考文献出现在测量和(或)设计语境下,相似的,讲述不同分析思路时,也会给出讨论测量和(或)设计问题的有关文献。

为了达到相同的目的,我们也通过在不同语境下重复使用相同的示例性实例。这样,在讨论测量问题(例如,题器内容)时所采用的一个实例,会在讨论设计问题(例如,内在效度)和(或)特定分析方法(因子分析)时,再次得到采用。

在学习一个主题时,大家是否以及何时精读本书的不同篇,这一决策取决于大家的背景、需求、目标和学习习惯等。因此,如果

大家的目标是得到一个主题的一般概念(例如,因子分析),而且大家能容忍由此而来的不可避免的模糊性,那么,大家只要阅读第4章我们有关这一主题的一般评注就够了。如果大家的目标是更好地掌握因子分析,或许是因为大家想更好地理解一篇应用因子分析的论文,那么,大家至少应该阅读第22章的一些章节。不过,如果大家想完全理解因子分析、培养应用因子分析的能力,那么,或许大家会发现,学习第22章和第23章,研读那里讨论细节和扩展的参考文献,都变成了基本要求。

### 参考文献

大家很快将会发现,大量的参考文献伴随着每一个主题。提供参考文献,是考虑到大家在需要它们的时候,能够很方便地得到它们。在从一种视角来阅读一个主题时(例如,获得一个主题的大体印象),大家很可能会忽略一些或全部参考文献;但从另一个视角来阅读相同的主题(例如,想要深化或扩展大家的理解)时,精读大多数参考文献就变成一种必需。

### 学习建议

和分析方法一起,我们给一些章节提供了学习建议。在前面,我们已经说明了动手进行分析的重要性,不能停留在只读不练的阶段。正是考虑到这一点,我们才提出了学习建议。我们给出了答案,以便让大家检查自己的作业。

在一些学习建议中,我们建议大家阅读引用的研究报告,再分析报告中的数据。这样做有几个理由:①我们认为,这样做可以让大家能够更好地欣赏报告作者的工作,更好地评估报告的发现、结论和意义;②看起来,不同的分析方法显得更“恰当”,或者,进行另一种分析方法的做法会强化文中所表述的观点;③我们相信,这将提高大家敏锐地阅读和评估研究的能力。

## 取 向

我们认识到,我们取向的一些方面会让一些人(或许也包括大家)感到吃惊,他们认为,这种取向过于粗糙、过度悲观。因此,我们要解释一下那些可能会留下这种印象的方面。人们可以形成一

个相反的印象：我们的目标不是散布绝望，而是知会大家在进行社会行为研究中可能遭遇的严重问题和困难。我们认识到，即使在认同我们的总体观点和意图的人当中，也有一些人会质疑我们在一个导论中讨论它们是否明智，因为这有可能让幻想破灭，让学生感到灰心。我们的意见是，学生不必受到保护，避免接触研究界的艰苦现实。相反，我们相信学生们会反感这种保护，特别是当他们发现"真实世界"的研究和教科书、课堂对它们的描述之间毫无相似的时候。

这浓缩成两个问题：什么时候，由谁来告诉学生研究过程中所伴随的复杂性？即使在学生准备不足的时候，也鼓励（实际上是期望）他们进行研究，这种倾向的背后是边做边学的信念。我们认定，在这种情形下，学生们最有可能学到的东西是，做科学研究几乎不需要准备什么，甚至不需要知识。

## 研究和测量的批判

和前面的立场一致，原草稿中的一个主要方面是对一些研究的批判性评估，这些研究采用了有关的分析和（或）测量方法。我们认为，这是集中讨论特定的错误应用和错误阐释的有用手段，因而会带来对有关方法以及总体研究过程的更好理解。因此，在大幅压缩手稿，让它具有合适的厚度时，我们十分犹豫地删除了大多数研究示例。即便是残留下来的极少数示例，也不过是原手稿中的影子。我们告诉大家这一点，是希望能够解释为什么我们只举了一些例子，而在讨论示例的时候，又倾向于一带而过。

需要注意的是，我们的本意不是总结或综述我们所评注的研究。相反，我们评述一个研究或测量的特定方面，目的是突出有关主题的要点，或者是一种错误的示例，我们希望大家能从中学习经验。正如我们在不同地方（例如，第 11 章）不断提醒大家的是，没有什么可以替代阅读一个研究的原始报告。

我们相信，反对批判研究同事的规范十分强大。为了避免离题万里，我们将简述一下这种规范（在我们看来）的来源。

首先，有人认为，批评同事会降低专业在公众眼中的公信力，更不必说资助机构。在我们看来，社会行为科学家不愿采取更批判的立场来对待已发表的研究，特别是对待建立在高度可疑的"发现"之上的失控主张，而恰恰是这种普遍趋势，才是造成社会行为

科学公信力消失的原因。

这样,才出现"建设性"的概念,批判必须伴随替代方案(当然是"更好"的)的建议。在我们看来,这种观念的根源可以追溯到一种态度(其他因素除外),即不愿承认在目前的知识和方法现状下,有些事情是不可为的。实际上,"只要观察和测量是可能的,'不可为'的概念就不存在"(Lieberson, 1985: 7)。

弗里德曼(D. A. Freedman)告诫我们应"启动一个新趋势"(Freedman, 1987b: 213),承认我们不知道如何进行一些研究。作为对这样的告诫的回应,我们相信,还有必要承认:①一些问题还无法得到答案;②缺乏恰当的条件和工具(例如,背景、测量、分析方法)时,有些事情是不可为的;③尽管是文化神话,付出的努力未必有所回报。总而言之,重要的是认识到,"有聊胜于无"这句话并不总是正确的。

我们急于补充说,我们并不是在布道完美。至纯乃上善之敌,这句格言也适用于科学研究的情形。我们并不否认,痴迷于完美(即绝不能出错的担心)可能会让研究陷于瘫痪的状态。但是,认识到完美不可及,并没有给任意胡为发放执照。在围绕着客观性的争论中,格尔茨(Clifford Geertz)也提出了一个类似的论点:

> 在这些方面,因为完全的客观性是不可能的(当然,的确如此),所以,人们可以让自己的情感信马由缰,这种论断从不会给我留下印象。诚如索罗(Robert Solow)所言,一个相似的说法是,由于完全无菌的环境是不可能出现的,因此,人们也就可以在下水道里做手术。(Geertz, 1973: 30)

假如我们所暗指的这种错误概念和错误应用相对比较罕见的话,我们也就没有理由在我们类似的书籍中这样做。正是因为它们的流行及其负面效应,我们才不得不直面它们。邓肯以最有力、最直观的方式表达了相似的关注:

> 如果存在一个清晰可见的社会科学领域,它清楚地认识到这些谬误,并树立一个禁止入内的告示,那么,我们也就没有必要遗憾这些有关研究阵营的漫画。但在我的专业内,恰恰不是这样。杰出的单篇论文和统计占卜的透明练习肩并肩地发表在一起。如果垃圾仅淹没脚面,那么,我们还可以趟过去。但如果垃圾淹没髋部,那么,最机灵、最讲究的工人也难

以避免让其产品沾上垃圾。引用记录这种判断的各种示例,将是令人生厌、冗长乏味的事情。(Duncan, 1984: 226-227)

当"最机灵、最讲究"的研究者都受到拙劣的研究实践的负面影响时,我们可以想象,它削弱新手和学生能力的巨大效应。

## 评审过程

在本章的前面章节,我们不止一次暗指专业杂志接纳或拒绝论文的评审过程。我们的评述不应当看做是加入到围绕评审过程的争论之中的努力[奥尔特曼(Lawrence K. Altman)关于"未曾召开的评审过程的会议"的新闻报道,参见 Altman,1989],我们也不想提出其他的战略和政策。我们的目标比较温和,只是警告大家,在同行评审的王国中,并不是一切都运作良好,因此,培养大家对阅读的论文保持健康的怀疑态度。自不必说,这样的一种取向,如果不是扎根于知识,借助于阅读时的清晰思考,那便是毫无意义的,甚至是有害的。

我们应当强调,现存体制困扰我们的是,一个具有基本知识的人粗粗一读,就能甄别出来的重大错误和严重误解,似乎躲开了"专家"评审者和编辑的目光。令人遗憾的是,在我们看来,这种事情发生的频率足以让人警觉。

总之,我们坚信,如果我们给大家呈现一幅社会行为科学过分乐观的景象,那么,我们就在做一件对大家有害的事情。更重要的是,我们衷心地希望,这样的取向将有助于大家成长为一个研究的敏锐的读者,成长为一个干练的研究者。

# 2

# 测量和科学探索

测量几乎遍及我们生活和日常活动的每一个方面。我们测量各种事物(例如,重量、温度、烹调成分、时间、距离)。时不时地,我们也被其他人(例如,医生、教师、上级、招生官员、心理学家)从不同的方面(例如,血压、成绩、产量、能力、态度、焦虑)进行测量。简而言之,我们所做的大部分事情、我们所作的决策以及与我们有关的决策,都包含这一种或另一种测量。

测量在我们的生活以及我们所生存的社会中所起的作用,在很大程度上,预测了各种活动和决策的有序性和可预见性,也就是预测了社会的秩序和运行;而我们几乎是机械地介入这些活动和决策之中(测量的历史回顾,参见 DuBois, 1970;Wainer,1987)。

其他除外,单单是监管机构和部门(例如,美国标准局)的存在,就足以证明测量在现代社会中所起的重要作用。它们的任务是设立各种计量的标准并监测这些计量。实际上,“针对几乎每一种物理的、化学的或生物的现象,都存在一个联邦政府强制的测量方法”(Hunter,1980: 869)。

尽管(或者正由于)测量很常见,但在不同的情境下、针对不同的人而言,测量的含义完全不同。除了意义的多样性之外,测量也不是目的,而是描述、区别、解释、预测、诊断、决策等过程中的手段。科学家和外行人似乎都赞同,无论如何界定科学(参见第7章),如果没有某种测量的话,科学是不可想象的(参阅 Brodbeck,1968,第七部分;Campbell,1952:第6章;Churchman & Ratoosh,1959;Feigl & Brodbeck,1953;Kaplan,1964:第5部分;Nagel,1931)。而且,在很大程度上,所使用的测量程序可以预测

科学的发展进程。玛吉诺(Margenau,1950: 369)曾深刻剖析这个观点,他认为,测量是"科学家对自然的最终诉求",它矗立在"理论与经验的关键节点上……是理性与自然的接触"(Margenau,1959: 163-164)。

下面,我们将首先考察测量的定义及优点,然后再讲解两个主要问题:测量尺度和测量与统计学的关系。

## 测量的定义及优点

和任何概念出现的情形一样,如果不加以界定,或者仅仅提供参考文献而没有指明一个特定定义时,测量的含义就一定会被模棱两可、混淆不清以及意见不一所包围。在行为科学关于测量的各种定义中,史蒂文斯(Stevens,1951,1959,1968)提出并细化的定义较为突出,尽管它没有得到广泛的接纳。他改进并拓展了坎贝尔(Campbell,1928)的表述,将测量定义为"依据规则或习惯把数字分配给对象或事件的不同方面"(Stevens,1951: 850)。

请注意,数字①是赋值给对象的不同方面,而不是对象本身。这样,举例来说,我们可以测量一个盒子的长度、宽度、重量、体积、颜色等,而不是盒子本身。同理,我们可以测量一个小孩的身高、体重、智力、焦虑、动机等,而不是孩子本身。

测量是针对人或物的不同方面,这种认识让一些常见的论证变得不相干,它们认为,在社会行为科学中,由于人类的独特性和巨大复杂性,测量要么是不可能的,要么是无意义的(最佳的情形下)。当测量各种对象的一个既定方面时,我们就会忽略它们在其他各方面的差异。这样,当我们测量对象的重量时,就会忽略其他各方面(例如,大小、颜色、形状)的差异,因为我们假定它们与当前的任务无关。出于某种目的,可能只有对象的重量是相干的;对其他目的而言,我们可能必须关注并测量其他方面。尝试解释人类行为的研究具有一个特征,即复杂性,其他部分来自我们必须关注并测量大量的潜在属性。

---

① 史蒂文斯指出,有些学者对"numerals(数词)"与"numbers(数字)"作了区分,并讨论了与这个区分有关的一些歧义(Stevens,1959: 19)。不过,他承认,在这方面,他自己也没有做到首尾一致(Stevens,1951: 22)。

与测量相干的事物，只能在一种关于待研究现象的、隐含或明晰的理论中加以决定。因此，测量暗含一种理论，它涉及与待研究现象有关的一组变量的操作或关系。也就是说，如果不存在一种智力理论，其他除外，它至少能够阐述智力与其他建构和变量之间的关系，那么测量智力的尝试就毫无意义。例如，在智力的多维结构论的语境下，吉尔福德（J. P. Guilford）拒绝了“单一智力的教条”（Guilford，1967：27）。而且，正是在自己的理论语境下，他才开发了智力运行的各种测量。理论不仅决定了测量哪些属性或方面，并且决定了如何测量它们。换句话说，理论界定了不同方面，随后才出现测量操作（参见第8、9章）。

## 同 构

就所测量的方面而言，赋值给对象的数字反映了对象之间的关系，这是测量的精要。这个观念（称作“同构”）意味着：两个组的元素之间存在一一对应①。

同构的一个原型是一幅地图与它所描绘的地理区域之间的关系。例如说，小镇和地图上用来表示它们的点之间，存在一一对应，这样，地图上各点之间的关系（例如，距离）也就反映了它们所表示的地理位置之间的关系。地图的有用性和方便性由此而来。测量就是将一组对象映射到一组数字上，这样，被测量的对象与赋值给它们的数字之间，就形成了同构关系。一个很明显的例子是测量重量，就是给一组对象中的每一个赋值一个数字，让数字之间的关系反映对象之间的重量关系（例如，一个对象是另一个对象的两倍重）。

## 测量的优点

就一个既定方面，对一组对象进行表述或区分的方法有很多。和其他方法相比，或许我们就可以看出测量的各种优点。我们可以尝试用文字来描述一组对象的重量（例如，重、很重、极重、不重、轻、较轻、很轻），也可以用一组数值来描述相同对象的重量，对照这两者，我们就可以看出前者的局限、歧义和潜在不一致。

---

① 有关同构的更精确定义及讨论，参见布罗德贝克（Brodbeck，1959）、科恩和内格尔（Cohen & Nagel，1934：137-141）和史蒂文斯（Stevens，1951：23）。

使用测量的一个巨大优势在于,我们可以应用数学这个强大的工具,来研究现象。一个数字集和一个对象集的各方面同构,对数字集的运算就可以让我们形成有关现象的规则性或规律的简明且精确的命题,如果没有测量所带来的优势,我们就无法达到这样的程度。例如,假如我们想研究和描述智力与成就之间的关系。如果仅依靠观察和文字描述的话,我们可能只停留在比较笨拙且模糊的命题上(例如,和低智力的人相比,高智力的人一般会表现出更高的成就)。与之相对照的是,测量一个样本人群的智力和成就,我们可以计算出两个变量间关系的一个指数(例如,相关系数),从而清晰、简洁地指明这种关系的方向和强度,这是文字描述所无法完成的。更何况,此后,这个关系指数还可以用于各种目的(例如,判定智力与成就之间的关系在不同种族间是否存在差异以及存在多大差异),或者,与其他统计量一起,形成一个方程,用智力来预测成就。

测量建立了数字与对象各方面的一种匹配,因此,我们必须了解"什么跟什么匹配"(Stevens,1968: 854),然后,我们才能有意义地阐释数字,决定哪些数学运算可以有意义地应用到它们身上。归根结底,这是尺度类型的问题,是接下来我们所探讨的主题。

## 测量尺度

史蒂文斯提出了下面四种测量类型(也称做"测量层次"),从最粗糙到最精细、按升序排列,依次是:定类、定序、定距和定比。(Stevens,1951)这个分类后来有过扩展和精炼(例如,Coombs ,1953; Stevens, 1959),不过,它足以满足当前的目的。史蒂文斯根据保持守恒的转换类型来界定尺度类型,守恒意味着没有方差、没有变化。贝尔(E. T. Bell)调整了凯瑟(Keyser)的一个命题,以令人折服的方式传达了守恒的观念:

> 守恒是变化中的不变性、一个流动世界中的淡定、各种构型的坚固,在无数不同寻常的转换的旋涡和压力面前,这些构型依然故我。(Bell,1945:420)

在尺度类型的语境下,守恒是指可以应用于数字的各种转换,它们不会改变数字所指征的经验关系的意义或阐释。在下面对四

种尺度类型的分别讲解中,我们将讨论并举例说明这些转换。

### 定类尺度

定类尺度就是把数字赋值看做是对象或对象类别的标签。换句话说,我们用数字来代替名称或其他任何符号,来鉴别对象或对象类别。[①] 定类尺度的最基本的例子,是将身份号码赋值给一组对象。身份号码这个惯例出现在各种不同的情境和背景下,如此显而易见,以至于无需再费笔墨,或者再寻找理据。不过,在测量的参照系下,定类尺度的重要性不在于它把数字赋值给单个对象,而在于赋值给对象类别。[②]

在我们的日常生活与活动中,分类和分类表起着重要的作用。如果不是经常(几乎是不由自主地)借助于分类,我们面对和处置冲击我们的大量刺激的能力,我们弄清楚原本是对象或事件的混沌旋涡的能力,都是无法想象的。我们使用各种规则和标准,将人、物、工作、科学专业、植物、事件或我们所拥有的东西进行分类。有些分类显得简单、"明显",甚至"自然";有些分类十分繁杂、极端复杂、完全不明显,因而造成意见分歧,甚至引发剧烈争论。

例如,当我们依据性别对人进行分类时,规则就很简单、自然、明显。相比之下,上帝让基甸把自己的部下分成两类的规则是:他想送回家的人和他带着去攻打米甸人的人。上帝让基甸把众人带到水边,然后,根据他们喝水的方式,对他们进行分类:

> 凡用舌头舔水,像狗舔的,要使他单站在一处。凡跪下喝水的,也要使他单站在一处。……耶和华对基甸说,我要用这舔水的三百人拯救你们,将米甸人交在你手中。其余的人都可以各归各处去。(旧约·士师记,7:5-7)

这是一个简单有余,但"明显"不足的规则。可以推定,自觉跪下的动作意味着起着偶像膜拜的一个指标的作用,当然,也可能存在其他的阐释。事实上,各种指标都可能从属于不同的阐释,这是指标的属性。我们将在后面的章节介绍有关指标与潜变量之间关系的概念。

---

① 史蒂文斯更喜欢将它们称做"numerals(数词)",因为"numbers(数字)"被用做标签。

② 一些研究者对两者进行了区别。史蒂文斯称前者是A型,后者是B型(Stevens,1951:25);纽纳利(Jum C. Nunnally)称前者为标签,后者为分类(Nunnally,1978:13-14)。

不论简单还是复杂,不论是广泛接纳还是激烈争议,分类都反映概念、变量(例如,社会经济地位、种族、党籍、宗教取向),并因此而成为隐含或外显参照系的一个组成部分。应用于科学研究时,分类是理论命题的组成部分。

有些研究者认为,分类不具有测量的地位。谈到测量的定类层次时,库姆斯(Coombs,1953:473)写道:“测量的这个层次过于原始,以至于人们并不总是把它看成是一种测量,但是,它是测量的所有较高层次的必要条件。”

为了满足定类尺度的要求,我们必须把对象分成一组互斥且穷尽的类别。也就是说,每一个对象只能被分到唯一的一个类别中,所有对象都能分类到所用的类别中。例如,将人们按照党籍进行分类,每个人只能是一个政党的党员,每个人都可以被分到所采用的类别之一。如果出现一些人不能契合当前类别中的任何一类的情形,那么我们就必须增加新类别,以满足穷尽性的要求。有时我们可以方便地采用诸如“其他”这样的一个类别,把那些不能契合前面类别的人归入其中。

不管分类的规则是什么,分到各个类别中的对象,不同的只是种类,而不是程度,我们只能这么处置它们。也就是说,定类尺度的类别没有次序。例如,在党籍上,民主党员不大于共和党员,反之亦然。他们之间彼此不同。分类的另一个属性是,分配到同组的对象,我们把它们看成是彼此相等的,除了界定类别所采用的差异之外,它们在其他方面的所有差异都不在考虑之列。例如,如果党籍的定义是基于一个既定政党的注册成员,那么所有已注册过的民主党员都应一视同仁,尽管他们可能在下列方面有所不同:对民主党目标的承诺、投票记录、缴纳党费和为党服务、性别、种族,以及其他任何可以想到的变量。

一个既定分类是否有意义或是否有用,如果我们不考虑最初使用这个分类的原因,不考虑是什么达成对这个变量的特定界定,我们无法回答这样的问题。分类是测量的一种形式。如前所述,测量是一种手段,而不是目的。只有在一个既定的理论或实践语境下,我们才能评估一个既定测量的意义和有用性。

不言而喻,不同的定义导致不同的分类规则,也可能导致对相同人、物等的不同分类。例如,依据种族对人进行分类,广义上依赖于种族的定义,狭义上依赖于所采用的类别。我们将在第8章中

讨论变量的定义问题，这里就不再详述。

不论二分或定类变量的定义如何，所采用的组别或类别就构成了一个定类尺度。对每个类别，我们可以赋值任何数字，只要不同的数字赋值给不同的类别即可。赋值给各个类别的数字，只不过是一个身份符号，因此，只要是一对一的替代，尺度守恒都能得到维持。也就是说，只要我们用不同的数字来标识不同的类别，那么我们就可以用其他数字替代任一数字。虽然我们可以自由选择任意一组数字来标识一个定类尺度的不同类别，但是，对特定目的来说，有些选择证明比其他选择更简便一些。例如，虽然任意两个数字（例如，1 和 22；0.06 和 23.73）都可以表示一个定类变量（例如，男性、女性；黑人、白人）的两个类别，但使用 1 和 0 来表示更简便些。在后面的章节（特别是在第 19 章）中，我们将说明，对编码二分或定类变量而言，有些数字更有用，因为它们有利于采用变量所进行的分析，也方便对分析结果进行更直接的阐释。

### 定序尺度

定序尺度是将数字赋值到人或物上，让数字来反映他们在一个我们感兴趣的属性上的排序。例如，如果我们认为人物甲比人物乙更友好（或更聪明、更好看），那么我们就可以将“2”赋值给人物甲，而将“1”赋值给人物乙。所赋值的数字并不反映在该属性上甲超过乙多少，而只反映两者之间“大于”或“多于”的关系，以“ > ”符号来表示。

在定序尺度中，对任意一对对象（如甲和乙）而言，如果甲大于乙，那么乙就不会大于甲，这种关系一定成立。这也称做“不对称”或“非对称”关系。当然，甲也有可能等于乙，形成对称关系。在这种情况下，我们可以赋值给甲和乙相同的数字，称之为“并列排序”。

在定序尺度上，对任意三个对象（如甲、乙和丙）而言，如果甲 > 乙，且乙 > 丙，那么甲 > 丙，这个关系一定成立。这也称做“传递性”。非对称关系并不一定是传递关系，如在国际象棋中，人物甲胜人物乙，人物乙胜人物丙，但是从这些条件中，我们并不能推论：人物甲能胜人物丙。

在定序尺度中，赋值给对象的数字只能反映“大于”关系，因此，对量表取值的任何单调转换，都可以保持守恒。单调转换就是

数字之间的排序不发生改变的运算。下面都是单调转换的例子:给所有数字加一个常数,对数字进行乘方运算,取数字的平方根,对数字乘以一个常数。

定序尺度所传递的信息比较粗略,也比较有限,因此,在各种转换所带来的反复无常面前,它可以经得起考验。下面的两个示例,可以说明定序尺度的各种局限,也可以说明对尺度数值的潜在误解。

假定有两组,每一组由 8 个人组成,这 8 个人按身高排列,结果如图 2.1 所示:线段上面的字母表示人,线段下面的数字表示按身高排列的顺序;(a)和(b)代表两个组。注意,在身高范围内,人们并非均匀分布。例如,在(a)组中,A 和 C 在身高上很接近,而 C 和 D 的差距相对较大;或者,D 和 B 的身高比(a)组中其他任何两个人的身高都接近。但是,当我们把身高的这些测量值(它们属于定比尺度,参见后面的讨论)转换成定序尺度时,这类信息便会丢失。例如,即使排序已知,在身高上,我们也不可能辨别 A(排序第 1)是和(a)组中排序第 2 的人(在这个示例中是 C)更接近,还是和(b)组中的人差距更大。

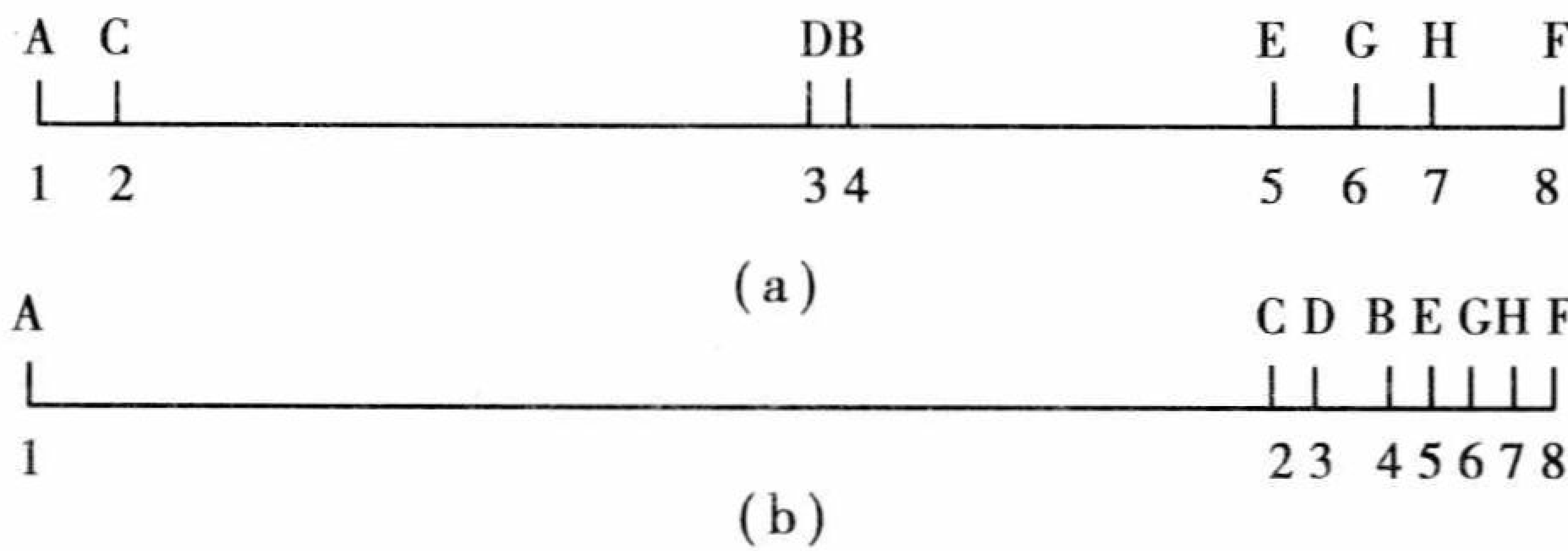



图 2.1

现在让我们来看一下两组的排列次序。很明显,赋值给不同组的次序,对它们进行比较,将毫无意义。在不同组中,两个人具有相同的次序,这并不意味着他们具有相同的身高。例如,(a)组中身高最矮的人,也有可能比(b)组中身高最高的人还要高。

现在,让我们利用图 2.1 来说明定序尺度适用的各种情形。这次,令 A 到 H 是 8 个对象(事件、人物等),(a)和(b)这两个排序是两个被访者对同一个对象的排序赋值。例如,A 到 H 是不同种类的食物,两个人给它们排序,来表明他们的偏好;或者 A 到 H 是不同的电视广告,两个人根据它们的效果来排序。就目前讨论的目的而言,假定字母 A 到 H 代表竞选职位的政客,两个人根据坦诚来

给他们排序。假如我们要求这两个人在“坦诚量表”上给这些政客评分,而不是给他们排序,那么这些字母就不会是均匀分布,反映出两个人给政客的评分不同。很明显,被访者给政客排序所得到的数字,并不能传递这类信息。

首先看一下(a)组的排序。假定给政客排序的人觉得,A 稍微比 C 更坦诚一些,而他们两人比其他人坦诚很多;D 稍微比 B 更坦诚一些,他们又比剩下的人坦诚很多,这个排序如图 2.1(a)所示。由于排序的限制,被访者并不能传递这些信息。由于自身的性质,排序只能传递相对意义上的强度,也就是说,被访者认为,A 比 C 更坦诚些,C 比 D 更坦诚些,等等。排序并不能告诉我们,在进行排序的人眼中,每个政客的坦诚程度有多大。因此,作为另一个例子,在图 2.1(b)中所显示的一个人的排序中,他可能觉得,政客 A 相对坦诚,而其余的人整体上都处在不坦诚的位置上。排序也没有提供这些信息。

注意,在(a)和(b)中,政客的次序相同。但是,在上述假定的基础上,很明显,得出“这两个排序反映相同的主观知觉的坦诚度”这个结论,就是一个错误。在某种“绝对”但未知的意义上,排序显示在(b)中的人甚至有可能认为,政客 A 的坦诚度低于政客 F,后者是排序显示在(a)中的人给予最低排序的政客。

上述讨论的要旨是,要求被访者进行排序的测量,只能用于研究个体之内的等级、偏好等。然而,它们不能用于个体之间的比较。卡特尔(Raymond B. Cattell,1944)提出一个术语“自比”,用来表示那些只能作个体之内阐释的测量,它不同于“规范”测量,即可以作个体之间阐释的测量。

## 定距尺度

当我们将数字赋值给对象时,除了满足定序层次的要求外,如果我们可以依据所测量的属性,对数字之间的差异进行有意义的阐释的话,那么我们就达到了定距层次的测量。换句话说,在定距尺度上,我们使用恒定的测量单位,这样,我们就能够有意义地表述对象之间的差异、比较这些差异,并将差异转换成比率。

定距尺度最常见的一个例子是温度的测量。例如,在摄氏尺度上,60 ℃不仅比 50 ℃高,而且是高 10 ℃。由于尺度的单位是恒定的,所以,下列说法也是正确的:60 ℃和 50 ℃之差等于 90 ℃和

80 ℃之差,或者说,60 ℃和 50 ℃之差是 37 ℃和32 ℃之差的两倍。

经下列线性转换,定距尺度保持守恒:

$$X' = a + bX$$

其中,$X'$是转换后的值,$a$ 和 $b$ 是常数,$X$ 是待转换的原值。用语言来表述就是:在定距尺度上,原值乘以一个常数 $b$,再加上一个常数 $a$,并不影响我们对该尺度上的差值或差值比率的阐释性质。众所周知,摄氏尺度上测量的温度,可以用下列公式转换为华氏温度:

$$F = 32 + 1.8C$$

其中,$F$ 和 $C$ 分别表示华氏度和摄氏度。

特别需要注意的是,定距尺度上的分值之差,表示为比率是有意义的,但这样来表示分值本身却是没有意义的。原因是,定距尺度上的零点是任意的,因此,这种尺度上的分值允许加上一个常数。把定距尺度上的分值表示为比率,这样的谬误可以用下面的示例加以说明:80 ℃是 40 ℃的两倍。这个命题是谬误,因为摄氏尺度中的零点是任意设定为水结冰的温度。零点的定义不同,这个比率也跟着变化。当我们把前面提到的温度转换到华氏温度上时,这就变得很明显,如下所示:

$$\frac{80\ ℃}{40\ ℃} \neq \frac{176\ °F}{104\ °F}$$

如果我们牢记定距尺度的这个特性,就相当于有了一个保险,让我们不会犯下面例子中的错误,这是爱丽丝和红后的一段对话(转引自 Carroll,1960:222-223):

> "五个夜晚比一个夜晚更暖和些吗?"爱丽丝壮着胆子问道。
>
> "当然,五倍的暖和。"
>
> "按同样的推理,那也会五倍的寒冷呀。"
>
> "正是呀!"红后喊了起来,"五倍的暖和,五倍的寒冷——就像我有五倍于你的财富,五倍于你的聪明一样!"

我们来看一个社会行为测量的例子,设想:(a)在智力的定距量表上,个体甲的得分是 120,个体乙的得分是 60。智力量表的零点肯定是任意的(我们如何定义智力为零?是绝对的意义?还是

以死亡为准?),因此,由此得出结论说:个体甲是个体乙智力的两倍,就出错了。(b)在社会研究的成就定距量表中,个体甲答对了60道多选题,个体乙答对了15道。虽然甲答对的题数是乙的四倍,但这并不意味着,在社会研究中,甲的知识是乙的四倍。如果我们想让这个结论正确,那么我们就必须证明,在社会研究中,测量中的零分(即没有一题答对)表示知识为零。正如社会行为科学中的大多数测量一样,当一个测量中的题器意味着表示一个领域,一个几乎是无穷大的领域时,毋庸置疑,这样的证明是不可能的。

## 定比尺度

在满足定距尺度的前提下,如果我们能决定一个真正的(或绝对的)零点,那么我们就达到测量的定比层次。即零意味着所测量的属性没有量。"比率"这个术语是指这样一个事实:在定比尺度中,任意两个分值的比率独立于尺度的单位。换句话说,当任意两个分值都乘以一个常数,即改变尺度的单位,它们的比率保持不变。定比尺度的最好示例是重量和高度的测量。例如,测量重量时,说重30磅的物体是重10磅物体的3倍,这样的说法是有意义的,并且,改用其他单位(例如,盎司、千克)来表示时,这个比率也不会发生变化。再例如,一个身高6英尺的人是身高5英尺的人的1.2倍,改变长度单位(例如,英寸、码、米),这个比率依然保持不变。由前面的讨论可知,在下列转换下,定比尺度保持守恒:

$$X' = aX$$

其中,$X'$是转换后的分值,$a$是一个正常数,$X$是转换前的分值。注意,只有这种转换才不会改变零点,才会消除负值。例如,加上一个常数,就会让零点产生位移,从而改变转换后的分值的比率。因此,重180磅的人是重90磅的人的两倍,当这两个分值都乘以一个常数,转换后重量的比率不会发生变化。但是,如果两个分值都加上一个常数,如40,那么它们的比率就会发生变化。即:

$$\frac{180}{90} \neq \frac{180+40}{90+40}$$

尽管我们听说过定比尺度,但在社会行为科学中,我们并不会经常遇到它们。对反应时间(即对知觉—运动任务的反应)的测量是心理研究中所使用的定比尺度的一个例子。

## 测量的指标和层次

大多数测量是间接的。也就是说,我们无法直接测量自己所感兴趣的现象,而只能从假定它所影响的或与它相关的一个指标中推导出来。用水银柱的膨胀作为温度的一个指标,就是前者的一个示例。在社会行为研究中,各种指标广泛地应用于各种建构(例如,动机、攻击性、工作满意度)的测量中。正是考虑到这样的应用,我们必须等到讲解建构效度(第 4 章)时,才能对指标进行详细讨论。

在当前的语境下,我们将只演示一个概念的测量层次的决定因素,不是我们当做其指标的事物的单位,而是这个指标与这个建构之间关系的性质。例如,一个指标的单位可能构成一个定距(甚至是定比)尺度,但是,当我们把这个指标当做一个特定概念的测量时,它的单位可能只是一个定序尺度。卡特(Lewis F. Carter, 1971:14)对这个问题的讨论非常有见地,他把收入与受教育程度作为测量社会地位的一个例子:

> 收入和受教育程度,每一个都肯定是对某物的保守测量(实际上是定比测量)。收入就是所报告的收入的一个测量,受教育程度就是所报告的在校读书年限的一个测量。但是,这两者如何与社会地位联系在一起?

在把收入和受教育程度作为社会地位指标的大多数应用中,卡特详细阐述了其背后难以成立的一个假定:即社会地位与前者中的每一个都是线性关系。这个假定就相当于说,等量的收入之差反映了等量的社会地位之差。不过,以卡特的例子为例,收入分别为 1 万美元和 1 000 美元的两个人,他们之间的社会地位之差,与另外收入分别为 21 万美元和 20.1 万美元的两个人之间的社会地位之差相比,根本不具可比性。如果我们考虑到受教育年限在下列点上的一年之差:(a)10 年与 11 年之间;(b)11 年与 12 年之间(后者意味着从高中毕业);(c)15 年与 16 年之间(后者意味着从大学毕业),它们所造成的结果潜在地具有相当的差异(例如,就业概率、工作类型、收入水平),那么,就受教育年限而言,不具可比性也同样成立。

总之,我们对测量层次的讨论仅局限在和分值阐释的相干性

上。这里,我们并不讨论如何决定一个既定流派的测量层次所涉及的一些重要问题和步骤,因为这需要我们详细讨论测量理论和量表模型。这个领域内的文献,数量上汗牛充栋,数学上比较复杂。此外,一些模型和程序也被开发出来,它们试图把定序反应转换为定距尺度,或将定距尺度转换为定比尺度。导论性质的文献,参见:Allen & Yen, 1979:第 8 章;Anderson et al, 1983;Nunnally, 1978:第 2 章。下面是一些书籍,专门一般性地介绍量表模型和技术,或者是它们在某个特定领域(例如,态度测量)中的应用:Andrich, 1988;Bock & Jones, 1968;Coombs, 1964;Dunn-Rankin, 1983; Dawes, 1972; Edwards, 1957b; Kruskal & Wish, 1978; Maranell, 1974b;Torgerson, 1958;以及 van der Ven, 1980。(同时参见本书第 6 章及其参考文献)

## 测量与统计学

不了解测量和统计学的人容易混淆二者或将二者等而视之。对那些讨厌公式、方程式、数字的人而言,尤其如此。熬完一门统计学课程之后,他们既不会认识到上测量课的需要,也不会坚持上一门测量课,他们把测量课看成是苦差事的别名。这种态度部分解释了人们并不理解,并几乎完全忽视测量在科学研究中的作用,这表现在社会行为科学中的许多学生和专业人士身上。

下面,我们将探讨测量与统计学之间关系的两个方面:(a)测量作为统计分析中所用数字的一个来源;(b)测量层次与统计分析方法。

### 作为数字来源的测量

简而言之,测量提供统计分析中所使用的数字。一个研究者测量一个或多个变量(例如,智力、社会经济地位、种族、性别),然后,在简单或复杂的统计分析中,使用所得的数字来描述或总结现象、估计参数、检验有关所研究现象的假设。因此,测量的性质和质量会影响统计分析的各个方面(例如,参数的估计是否有偏差以及偏差的程度),这应当不出所料。总之,一个统计分析所得到的各种结果,其阐释和意义离不开测量的属性,正是这些测量在第一场合生成了数字。

虽然是老生常谈,但似乎仍有必要提醒研究者和研究结果的使用者,当用于统计分析的一组数据没有意义时,所得结果也没有意义。忽视这个老生常谈所带来的危险越来越大,这是越来越多使用复杂分析技术(例如,因子分析、判别分析、多元方差分析)的后果,因为只要具有使用各种计算机程序、进行数据分析的必要基本技能,人人都很容易使用这些分析技术。面对充满各种指数和统计显著性检验的一大叠输出纸张,人们很容易忘记输入电脑的数字的意义。①

令人遗憾的是,在大多数社会行为研究的基础上,我们可能会形成这样的印象:不管数字如何获得,不管数字意味着什么,它们都是送往统计学磨坊中的原料。而且,它们甚至还可能让我们相信:只有使用统计学,我们就可以不知不觉地将无意义的数字转换为有意义的数字,分析越复杂、越老练,事情就注定变得越有意义。这种取向所带来的有害结果不计其数。研究者可以通过使用各种复杂的统计分析,隐瞒(甚至不留痕迹地掩盖)数据中的各种缺陷。如此取得的"结果"可能与一个研究者所声称的研究问题,几乎(或完全)没有任何关系,尽管对研究者本人和研究报告的读者来说,这一点可能并不明显。

我们相信,大多数研究者并不会有意欺骗读者,但是,他们会深陷统计学的杂耍中,甚至达到自欺的程度。之所以会这样,是因为他们对测量在研究中的角色缺乏理解,或者是因为他们对统计分析的"魔力"近乎于迷信。尽管如此,由于进行最复杂统计分析的计算机程序变得很普及,对统计分析的误用也就随之急速增加,这常常形成多个层面,其下面是质量可疑(甚至是无意义)的数据。研究者和读者们游荡在一个虚幻的世界,其中住着因子、成分、负荷、模式、方程、函数等,给人留下这样的印象,常常是不可避免的。

以总和指数为基础的数据分析,其研究报告的问题尤其严重,因为进入这个指数的每一个元素的信息,我们几乎或完全得不到,因而就无法评估这个总和指数的质量。一个常见的情形是,我们对多题器测量进行加总分,然后用在复杂分析(例如,因子分析、结构方程模型,参见第 22—24 章)中,但却没有在第一场合提供必要的信息,以评价总分的价值或意义。不检查自己是否正在把苹果

① 我们将在第 16 章讨论这个问题及相关问题。

和橘子相加，就把各个题器的分数加总，令人遗憾的是，这个倾向相当普遍，并且，这是我们先前所提到的对测量问题缺乏注意和关注所带来的一个后果。

上述讨论的目的是要引起大家注意，在研究中，仔细审查所使用的测量十分重要。若有必要，我们还将在本书的各个部分，讨论测量的特定方面（例如，效度、信度）对统计分析的效应。

## 测量的层次和分析方法

现在，我们将从另一个角度来考察测量和统计学，在社会行为科学中，这个角度在心理测量学家、统计学家和关注研究设计与分析的研究者中，引发很多争论，而且常常带着情绪。这个争论可以追溯到史蒂文斯（Stevens，1951）的著作，在这本书中，他依据测量所采用的测量层次（参见前文），提出了一个"许可的统计量"（Stevens，1951：25）分类。举个例子来说，史蒂文斯主张，对定序层次的测量，我们不应当计算均值和标准差。

关于统计学与测量层次之间关系的文献非常多，有些学者强烈支持并阐述史蒂文斯的立场，有些学者则反对其立场，只是冷嘲热讽的程度不同。伯克（Burke，1963）对这两大阵营（即"测量导向"和"测量独立"）的立场作了很好的概括和讨论，下面所举的例子，不过是在这个问题上的几次交锋而已，我们的目的在于传递它在一些论辩者身上所唤起的情绪反应（最近一次反驳史蒂文斯批评者的尝试，参见 Stine，1989）。

在讨论关于"依赖尺度的错误"之前，沃林斯（Wolins，1982：29）写道："著名的心理学家史蒂文斯（Stevens，1951），把'测量'这个概念从物理学中连根拔出，然后扔到了心理学中。没有根，它就腐烂了，恶臭满天。直到现在我们还在扫清这个垃圾呢。"有一本论述测量和统计学的书，按照史蒂文斯的尺度分类和统计学的结构编排。在评述这本书时，凯泽（Kaiser，1960b：413）的结论是："这是一本错误连篇、粗制滥造的书……混杂着对史蒂文斯测量尺度的幼稚盲从，对现代统计学理论，则明显表现出几乎是完全的无知。"

洛德（Lord，1953）用讽刺故事的形式嘲讽"测量导向"的立场。简单来说，他讲了这样一个故事：一位教授习惯于对定序层次上的分值进行均值和标准差运算，他对这种做法深感内疚，以至于神经崩

溃,被迫退休。为了酬谢他以前的工作,学校特许这位教授卖"橄榄球号码",并提供给他一批球衣号和一部自动售货机。一切都很顺利,直到有人明显改动了这部自动售货机,导致新生队的号码售价太低,并遭到新生们的抗议。为了调查出事原因,这位教授求助于一位统计学家,这位统计学家不费吹灰之力,就对号码进行了各种计算,包括均值和标准差。这位教授感到吃惊,抗议道:这些橄榄球号码甚至都不能构成定序尺度。这位统计学家答道:"号码自己并不知道啊。"(Lord,1953: 751),他继续解释道:"因为号码不记得自己从哪里来,所以,无论如何,它们总是以相同的方式行事。"(Lord,1953: 751)或许大家已经猜到,这位统计学家的一席话,让这位教授完全康复,他又开始了自己的教学生涯,对学生分数的计算均值和标准差,也不再有丝毫的迟疑,更不用提内疚感了。

我们复述这个故事,是因为上述引文中的这位统计学家的做法,几乎已经成为"测量独立"立场的战斗口号,其形式一般为"数字本身并不知道它们来自哪里"(参见 Gaito,1980: 564)。当然,这句话的真实性是不可否认的。但是,这并不意味着阐释数字的人就没有了解数字来自何处的责任。下面的这个相关的轶事,是斯坦普(Stamp,1929:258-259)所举的一个示例:

> 考克斯(Harold Cox)讲了他自己年轻时在印度的一个故事。他向一个法官,一个英国人,同时也是他的好友,引用了一些统计值。他的朋友说:"考克斯,当你岁数大一些的时候,你就不会如此肯定地引用印度的统计值。印度政府非常热衷于搜罗统计值。他们收集统计值、汇总、取 n 次方、求立方根,然后,拼凑成精美的图表。但是,你永远不要忘记,这些数字中的每一个,最初都来自乡村更夫(chowty dar),他想填什么,就会记什么。"

有趣的是,"测量独立"的研究者们在各种场合总会引用洛德的话,来支持自己的立场,但是,洛德对其原始命题的批评所作的回应,他们经常只字不提。究其原始命题,洛德(Lord,1954:265)写道:"如果这些文字让大家忽视了实际存在的严重错误,那将是可悲的。"他列举了各种需要和不需要等距假定的命题示例之后,洛德写道:

> 我们的结论是:在阐释定类和定序数字的算术运算的结果

时，我们必须付出十二分的小心。因此，在一些情形下，至少就检验零假设的目的而言，我们能够严谨、实用地阐释这些结果。

我们可以随心所欲地运算数字，但是，对于数字运算所得到的结果，我们的实质性阐释却取决于把数字赋值给对象时所附加的意义，即测量模型。我们非常赞同海斯（William L. Hays，1988：71-72）的观点，他说："只有统计结果的使用者们（研究者们和读者们）才能够判断：数字结果是否被重新阐释为有关事物属性的一个有效命题……作为一门学科，统计学在这个问题上相当中立。"

如先前所指出的，我们只能在一个既定的实质情境下，决定一个尺度的层次。一个尺度"不是一个计量局专员所能决定的事物，不是他能凭自己的感觉，正确决定一个尺度可以称做这件事物而不是另一个事物"（Cliff，1982：12）。就这个方面而言，统计分析的恰当性也不是由统计局专员决定的。

对我们正在讨论的问题，大多数研究者和读者很可能并不关注。我们相信，这将带来一个后果：对测量问题的不幸忽视，这是我们先前讨论过的。不过，也有相当一部分学者仔细、认真地思考了上述两种立场，得出一个结论：在社会行为科学中，严格固守任何一个立场，都得不到测量现状的支持，如果我们考虑到违反特定统计方法背后的各种假定所带来的各种后果时，这种固守也没有任何用处。我们相信，这种实用主义取向最为合理，并用关于这种取向的一些评注来结束这个章节。

首先，我们注意到，在史蒂文斯的晚期著作（Stevens，1968）中，他在一定程度上也支持实用主义取向。在"调和及新问题"的标题下，史蒂文斯探讨了这两种立场，并提到了"评估违规工资的实用主义难题"（Stevens，1968：851），他的结论是：

> 因此，问题应该转向：不恰当的统计量如何导致有偏结论以及有偏的程度；而不应当转向：测量尺度是否决定一种统计程序的选择……通过详细说明成本，我们可以将看似禁区的问题转变成可计算风险的问题。（Stevens，1968：852）

在社会行为研究中，所采用的大多数测量，究竟是在定序层次上，还是在定距层次上？这是有关测量和统计量争论的主要来源。实用主义者（例如，Borgatta，1968；Borgatta & Bohmstedt，1981；Gardner，1975；Labovitz，1967，1972；Nunnally，1978）强有力地论证

说,在社会行为研究中,所采用的大多数测量明显不在定距层次上,但严格来说,它们也不在定序层次上。换句话说,我们所用的大部分测量,并不局限于表示“大于”或“小于”,像一个定序尺度一样,而且还表示差异的程度,尽管这些程度可能不能用等距单位表示。基本的一些实例是成就、智力、态度等的累加测量,这些测量处于定距层次与定序层次之间,即所谓的“灰色”区(Gardner,1975:53),如果把它们当做定序层次上的测量进行处理那么就有可能导致严重的信息丢失。

违反统计背后的假设会带来一系列后果,从这个角度,一些学者探讨了测量与统计学之间的关系问题。例如,纽纳利(Nunnally,1978:17)论证并试图证明“在行为科学的大多数研究中,采用重视定距层次的数学和统计分析方法,并没有坏处”。拉博维茨(Labovitz, 1967,1970,1972)是这种取向的最直白的倡导者之一,他认为:

> 处理定序变量时,仿佛它们服从定距尺度,这得到经验证据的支持……把定序变量当做定距变量来处理,虽然会伴随一些小错误,但是,通过使用更强大、更敏感、更发达、阐释更为清晰的统计量,加上已知的抽样误差,就可以抵消这些错误。(Labovitz, 1970:515)

值得注意的是,拉博维茨和纽纳利所代表的立场,并不意味着“不管测量质量如何,一切都行”。因此,他们也有批评者(参阅Wilson,1971)。在这个问题上,不论我们的立场如何,高质量的测量无可替代,这是毋庸置疑的。提高测量质量,应该成为社会行为科学优先考虑的问题之一。

## 结束语

我们希望,这一章能够让大家清醒地认识到测量在科学探索中所担任的关键角色。这尤其重要,因为在阅读社会行为科学的研究报告时,我们所得到的印象与我们试图去建立的事物之间,可能会形成强烈的反差。正如第1章所指出的,在研究报告中,我们经常忽视、傲慢地(甚至不加思考地)处置测量问题。因为一些测量是“现成”的,因为别人用过这些测量,因为不存在“更好的”测

量，所以，我们就采用这些测量。有些研究者对研究的其他方面（例如，理论表述、设计、分析）展现出较强的怀疑、细心和熟练，但对充其量只能算是粗略的测量，却轻易地相信，这令我们不得不十分惊讶。

在各种领域研究中，许多学者都关注到测量的糟糕状态。在态度研究的广阔领域中，就很多社会行为科学家对测量问题的冷漠问题，马拉内尔（Maranell，l974a：xii）写道：

> 如果我们忽视测量问题，那么我们将遭遇的效应和结果，就像天文学家被迫使用破裂且没有校准的棱镜、测绘员被迫使用橡皮尺或干脆不用尺子、物理学家被迫使用忽快忽慢的手表一样。

在回顾了市场研究的"现状"之后，雅各比（Jacoby，1978：91）提出：

> 鉴别好的测量、清除差的测量，这样的尝试几乎看不见……大多数测量之所以是测量，仅仅是因为有人说它们是测量，而不是因为已经证明，它们符合标准的测量准则（效度、信度和灵敏度）。

在不同研究领域中，尽管所使用的测量质量各不相同，但上述观察一般适用于许多（即使不是大多数）社会行为科学的研究领域。

# 3

# 准则关联的效度

在广阔的测量理论和实践背景下，本章和下一章将探讨效度问题。我们首先将简要回顾效度的含义和定义，然后将用剩余的篇幅讨论准则关联的效度。在这部分，我们将首先讨论准则的含义和准则的不同类型，然后讲述预测，我们将特别强调预测效率、分组预测和选择偏差。

## 效度：含义和定义

即使是泛泛阅读关于测量、研究设计的书籍，或者是专业杂志的研究报告，也足以让我们明白，在不同的语境下，不同的作者以不同的含义使用“效度”这个术语。例如，在测量语境中的“效度”含义不同于它在研究设计语境中的含义（参见第 10 章）。而且，在每个不同的语境中，我们会碰到效度的不同定义和不同类型、种类的效度之间的区别。

有趣的是，负责为教育和心理测量提供标准的数个专业协会的联合委员会，却没有达成对“效度”的定义；而只是提供了或许可以称做“效度的特征描述”的东西：即效度“是指从测验分数中所作出的特定推论的适当性、意义和实用性，测验验证（validation）是不断积累证据来支持这些推论的过程”（American Psychological Association，1985：9）。

上面的表述所面临的一大难题是确定：什么才能构成“适当的”“有意义的”和“有用的”推论——这几乎是一个不可能完成的任务，别的除外，这要求界定这些意味深长的术语，描述实现它们的条件和途径。或许表达这项任务复杂性的最佳方法是指明，尽

管“测验分数”和“测验验证”是较狭义的术语，但联合委员会的这个表述仍可以应用到科学研究的广泛领域之上。事实上，正如本章和下一章所详述的那样，测量验证是科学研究的一个实例，包含着科学研究的全部。因此，在第二部分讨论科学和科学研究的大部分内容，对测量验证过程中所包含的因素具有直接的影响。

联合委员会的表述清楚地表明了，效度（或准确地说，验证）不是指测量本身而是指在测量所得分值的基础之上的推论。简言之，“我们不是验证一个测验，而是验证对来自特定程序的数据的一种阐释”（Cronbach，1971：447）。（关于验证过程的各个方面的专题讨论会，参见 Wainer & Braun，1988）

因此，随着研究目的、被访者和进行推论的情境不同，推论的效度（恰当性、意义、实用性）也会有所不同。例如，用一个既定的词汇测验的分数，来推论一个人的学习成绩，与预测他们在大学或工作中的表现相比，这个推论更有效度。预测的效度会随大学课程的种类或所考察的工作类型的不同而变化。很明显，相同的词汇测验的效度也随着被访者的年龄、民族、种族、教育背景等因素而有所不同，这仅仅只是列举了一些因素。为了把模糊性降低到最低程度，至少有必要明确说明，从一组分数中得出的推论是出于什么目的、对象是谁、在什么情境下。

虽然我们使用测量的目的千差万别，但是，把各种目的归类为几种类别，不仅有可能，而且有实用性。与测量的效度有关的，一个普遍使用的三分法是：（a）内容；（b）准则；（c）建构。本章和下一章所讲述的大部分都与这些术语在验证过程的情境下的含义有关。因此，这里我们仅指明它们的定义。“内容”是指内容的某些领域（例如，社会学习、词汇、工作表现）；“准则”是指某些结果（例如，高中毕业、旷课、青少年犯罪）；“建构”是指某些特质或属性（例如，智力、态度、动机）。

无论采纳哪一种分类，重要的是要牢记：效度是“单一概念”（American Psychological Association，1985：9）。因此，就编排和讨论目的而言，依据推论类型进行分类，虽然便捷，但这并不暗含着一组排他、穷尽的类别，更不暗含着不同类型的效度。

需要强调的最后一点和我们所看到的下列事实有关：直到最近，有关效度的讨论和报告都是依据类型而进行表达的，而且，上述三分法（内容、准则和建构）在一段时间内成为一个主导性分类。

(参见 American Psychological Association,1966,1974)有些学者并不赞同不同的"效度类型"概念,但他们还是在十分勉强地使用它们,因为这是"这个领域中人们的传统习惯"(Ghiselli et al,1981:267)。

邓尼特(Marvin D. Dunnette)和博尔曼强烈反对"效度类型"的概念,他们(Dunnette & Borman,1979:483)认为:"效度具有不同的类型,其含义会导致混淆和混淆之上的过度简化。"在一次关于效度的感性讨论中,盖恩(Guion,1980:386)宣称,我们把三种"类型"的效度看做是"类似三位一体(指圣父、圣子和圣灵)的事物,它代表着通向心理测量救赎的三条不同道路。如果我们无法证明一个类型的效度,我们还有另外两次机会!"

"准则"和"建构"这两个术语,其用法常常和"推论验证"和"证据验证"两个过程相联系,而不是用来表示"效度类型",因此,我们就冒着滑向后者(效度类型)的风险。[①] 尽管我们可以尝试预防这种风险,但我们认识到,我们也可能会失败。在我们的表述暗含着效度类型的情形下,我们希望大家明白:这并不是我们的本意。总之,我们相信,假定对验证过程的一个分类没有变成拜物教,假定我们还能够清楚地意识到,不同的目的只是一个相同过程的相互关联的方面,那么依据主要目的而对验证过程进行分类,就是便捷的。

我们并不情愿把效度分成两章来讲解,有了前面的铺垫,大家听到这句话,就不会感到惊讶了。归根结底,便捷性的考量胜出。我们奉劝大家将本章和下一章看成是同一单元的两个方面。

## 准 则

广义地说,一个准则是我们希望借助其他变量的信息来解释和预测的任何一个变量(例如,学习成绩、投票、敌意、生产力、吸毒、旷课、青少年犯罪)。科学哲学家曾涉猎过解释和预测的研究。有些人(例如,Hempel,1965)主张,在结构上和逻辑上,解释和预测同一;有些人(例如,Scriven,1959)则反驳道,它们是不同的运算。不论我们在这一问题上的立场如何,下列命题依然成立:总会存在

---

① 参见第4章,有关内容与"内容效度"的评述。

一些情形，其中，我们能够预测一个既定现象，却不能解释它，反之亦然。（较好的探讨，参见 Doby，1967：第 4 章；Kaplan，1964：第 IX 章）

在准则关联的效度中，预测是焦点，压倒一切的关注是一个准则的预测成功率，“是什么过程导致所预测现象的出现？”这个问题能否得到解释，则不在关注之列。为了强调这个论点，纽纳利指出：“这样，如果掷蹄铁的精度与大学里的成功高度相关的话，掷蹄铁就是预测大学里成功的一个有效测量。”（Nunnally，1978：88；同时参阅 Cook & Campbell，1979：296）

上述内容并不一定暗示：对预测感兴趣的一个研究者和实践者，必然对解释不感兴趣。正如卡普兰（Abraham Kaplan，1964：350）所指出的：“如果我们能够在一种解释的基础上成功进行预测，那么我们就有了很好的（或许是最好的）理由来接纳这种解释。”不过，预测现象是可能的，也具有实用性，但解释要么是不存在，要么是模糊不清；在欠发展的学科和应用情境中，这种情况经常发生。

为了预测目的而应用心理测量，我们可以把它们刻画为“心理学技术”（Loevinger，1957：636），它们和心理学理论形成对照。这样的应用示例不胜枚举。例如，在工作场合，如果能够在一群职业申请者当中，预测出谁具有成功的能力、谁具有出事故的倾向、谁会对工作更满意等，那么无论对雇员来说，还是对雇主来说，这件事都是有益、有用的，只是程度各异、理由不同。即使缺少对现象的解释或解释不清，这些预测也是有用的。在学术场合，如果在申请者中预测出谁有更大的概率在大学中表现良好，那么尽管我们还无法透彻地理解之所以这样做的原因，或者我们对原因还存在着争议，但这件事还是有用、有益的。用耳道绒毛和耳垂来预测冠状动脉疾病还是有价值的（参阅，Elliott，1983；Wagner et al，1984），尽管对前者和后者之间的关系，存在多种相互矛盾的解释。

我们应当牢记解释和预测的区分，以避免错误阐释研究结果。对一个为预测而设计的研究，如果我们把它的结果阐释为对所预测现象的解释，那就大错特错了。我们将在许多章节（例如，第 14 和 18 章）中讨论并举例说明这个论题。（参见 Pedhazur，1982，第 6—8 章）

总之，当我们的关注点是利用预测因子来实现预测准则的目

的时,准则关联的效度之问题和流派就变得十分重要。在讨论它们之前,我们有必要讨论这个准则的一些方面。

**准则的性质和种类**

一个特定准则的选择,很大程度上取决于作出选择的个体的价值观和目标。不管这个准则是制造车间的生产力、学习成绩、婚姻满意度、健康,还是对少数民族的态度,在一个既定情境下针对特定人群,能够决定这个准则是什么的人认为重要的事物,才是关键。

在我们的生活中,准则随处可见。当我们想到自己和他人是成功还是失败时,我们就是在公开或隐含地使用一个或一组准则。当我们想到成功完成一个既定任务,擅长一个既定工作时,这是最明显的示例。即使人们看起来是在共用一个准则,但他们经常会在其定义上出现意见分歧。因此,大多数人可能都会赞同:教师、律师、医生、法官、护士、售货员、卡车司机等,应当能够娴熟、成功、高效地胜任自己的工作;但对这些形容词的定义,他们可能会存在很大的分歧。

一方面,存在赞同准则的表象;另一方面,界定准则又存在困难。很可能正是因为如此,在试图预测准则时,有些个人或机构几乎没有给予它们应得的关注。芬彻(Fincher,1975:495)把有关准则恰当性的不加批判的假定,恰如其分地称做"无知无畏"。詹金斯(Jenkins,1946:93)也注意到"心理学家通常倾向于接纳这个默认假定:准则要么是上帝的赐予,要么是躺在那里等待被人发现的事物"。

令人遗憾的是,自从詹金斯提出"对什么有效"这个问题以来,情况并没有多大改观。研究者依旧倾向于使用模糊和总体的术语来设定一个准则,然后寻找一些预测因子,能够对这个准则进行最优的预测。这样,当我们认为研究没有达到预期时(这种情况经常发生),一般来说,受到指责的是预测因子。不过,能够确定的是,预测程序"只能和准则一样好"(Thorndike,1949: 119)。而且,"准则关联的效度的致命弱点当然是准则本身"(Linn,1984: 38)。

对一个准则的界定,试图达成一致意见,其中所遭遇的困难,可以用下面的事件作很好的说明。不久前,时任美国大法官的伯格(Warren E. Burger)声明,在美国,大约半数的诉讼律师不称职,

没有资格代表当事人;在律师界,这句话激起公愤(参见 *The New York Times*,1997 年 12 月 4 日)。在一篇题为《测量能力:论辩一个无法界定之物》的文章中,戈尔茨坦(Goldstein,1978)描述了一场辩论,它发生在美国律师公会的大会上,是对伯格指责的回应。在大会上出现一项动议,要求这位大法官要么拿出数据来支持自己的指责,要么收回成见。但这项动议并没有通过。可能的原因是,大多数的与会代表意识到,在对"能力"的量化界定无法达成共识的前提下,绝不可能收集到这样的数据。

律师们(或是任何工作任务复杂多样的专业成员)对"能力"的一个定义,无法达成共识,这并不出乎意料。例如,一个姓里夫金德(Rifkind)的律师曾经做过尝试,他想刻画一个伟大的诉讼律师。按照里夫金德的说法,一个伟大的诉讼律师必须"热爱工作、坚韧不拔、记忆超群、才思敏捷、洞察入微、口才雄辩、仪表堂堂和嗓音悦耳"(Goldstein,1978: E7)。

里夫金德的这些刻画,抓住了一个伟大诉讼律师的精髓。即使我们能够对此达成一致,但毫无疑问的是,关于这些要素之间的相对重要性一定存在相当程度的不同意见,至于每个要素的定义,意见分歧的程度甚至会更大。例如,"坚韧不拔""洞察入微"和"口才雄辩"是否同样重要? 进一步说,它们各自的定义是什么? 虽然里夫金德的表述带着美丽的光环,但其模糊性却排除了一个量化定义的可能。

其他专业中界定"能力""成功""绩效"的各种尝试,可以无数次地重复上述示例。这足以说明,罗兹布姆(Rozeboom,1966:194)为什么会把准则的定义看做是"一个乱人心智的谜团"。

由于界定和量化准则中的各种困难,一些研究者和机构常常诉诸"可预测的准则,而不是诉诸恰当的准则"(Wallace,1965:411)。例如,怎样才算是一个"好"学校? 教育家、政策制定者和一般公众都表现出极大的兴趣。在"学校绩效"的标题下,相同的问题有时也会被提出来。不管用什么词汇,虽然所有人似乎都同意:学校应该是好的或有效的,但是,对这些术语的定义,却无法达成一致;如何测量好的或有效的学校教育,就更无从谈起了。

关于学校,有些事物相对容易观察和量化,这样,常常是一种默认,它们就变成了"好"学校的一项或一组准则。例如,我们可以相对容易地观察到下面事物:一个学校中,科学实验室的数量、学

校图书馆中的藏书量、班级规模、每个学生的开支、教师工资、学校管理人员的数量和标准化成绩测验(SAT)的平均分数。

需要注意的是,前面只参考了变量的计数和量化的一些形式。在这些变量中,有些变量比其他变量更容易量化,因为它们能够接纳的界定范围,潜在比较窄。例如,对照一下图书馆藏书量和教师工资,很明显,后者比前者的界定范围更广。

假定我们并不满足于只是进行计数,也想就上述变量中的一些或全部提出一些问题,如质量或使用。这将带来相当的复杂性。例如,是什么成就了一个"好"科学实验室?如果分配到这里的教师没有资格正当地使用它,那么它的好又是什么?如果没有人使用它或没有充分使用它,那么一个实验室的好又是什么?还有,怎样才算是充分使用呢?

在识别好的或有效的学校的多年探索中,准则的变异和变迁,完全不亚于在其他社会领域中所发生的时尚和潮流。因此,在一些地方或一个时期,好学校的准则是学生们在一些认知变量(例如,学习成绩)上的表现;在另一些地方或另一个时期,这个准则可能就是一些情感变量(例如,动机、态度)所构成的指数。

很难就准则达成一致的另一个领域是治疗的效果。例如,马斯特斯(William Masters)和约翰逊(Virginia E. Johnson)因为没有说清楚性治疗的成功准则,所以备受批评。为了回应批评者,据报道,马斯特斯博士曾说:"我必须利用七周的时间来写完这本书,忘记把最终准则加上。"(Wolinsky,1983: 2)另外,据报道,在记者招待会上,马斯特斯又声明:

> 在《人类性失当》中所呈现的个案,所采用的准则实际上是:在所有性交机会中,一个女性至少需要有50%以上的机会达到性高潮,才算作是一次"成功"。
>
> 在性无能的治疗中,在所有性交机会中,75%以上的机会能够达到并保持勃起状态的能力,就被界定为"成功的治疗"。(Brody,1983b: A13)

不难想象,在总体上,专业人士和门外汉都会质疑这样的准则,也会质疑其特定方面(例如,对性高潮的定义和测量)。

## 作为准则的评级量表

界定和测量准则面临很多困难,因此,雇主和研究者经常使用

评级量表。建构几个评级量表，让管理者评估工人、学生评估教授、治疗师评估病人、病人评估治疗师等，还有比这更简单的方法吗？当我们把评级量表作为准则时，在非常严重的问题当中，我们经常会碰到的是：(a)没有定义或定义模糊；(b)对评级背后的知觉过程，没有给予充分的注意，或完全忽视；(c)把各种评级组合成一个总体指数或几个子指数的方式。我们将在第6章(参见评级量表部分)中，对这些问题及相关问题予以讨论。

## 终极准则和中间准则

综上所述，应当比较清楚的是，准则的定义和测量都是非常困难的，并且我们常常躲避不了这些困难。称为“终极准则”的事物，尤其如此。

终极准则。桑代克(Thorndike，1949：121)曾经对准则问题有过精彩的讨论，就终极准则，他这样表达：“终极准则之‘终极’，是指在判断结果时，我们不可能越过它，采用更高、更远的标准。”一个终极准则就是一个终极目标，被我们看重的事物就是这个目标自身，因此，预测它所必需的成本也是值得的。例如，预测谁会成为好(有效的、敏感的)医生(律师、飞行员、教师、秘书、管理者)，社会认为这很重要。

即使当我们对终极准则达成极少的一致意见时，我们也应当认识到，准则是多面的、动态的。这样，就一个既定层面而言、一个既定情境下、一个既定时点上，有可能被认为是表现良好(或令人满意)的事物，就另一个层面而言、另一个情境下、另一个时点上，则有可能被看做是差强人意的事物。[①] 可想而知，好医生的准则会有所变化，其他因素之外，它会随着从医学院毕业时间的长短、专业类型和背景的不同而不同。在大多数场合，一个终极准则是一个建构、一个抽象。因此，在下一章，我们对建构和建构验证的讨论，也同样适用于作为建构的准则。

因为准则一般是动态的建构，在一个既定时点、一个既定背景下，一个人可能被评为“身材高”，是“成功的”等；在另一个时点、另一个背景下，这个人又可能被评为“身材矮”，是“较不成功的”或“失败的”；这种情形并不会让我们感到奇怪。实际上，如果有可能

---

① 有关“工作绩效准则”的动态性质的讨论，参见吉塞利和海尔(Ghiselli & Haire，1960)。

的话,在一个终极准则上,对一个人地位的盖棺定论,可能是对其一生表现总和的一次评估。正如诗人奥维德(Ovid)所言:

> 死前无人配称幸福;
> 总要等到临终的那一天,
> 在他身后留下终审。
> (引自 Montaigne,1965:54)

中间准则。面对定义和测量终极准则所遭遇的各种困难,人们常常求助于所谓的“中间准则”(Cureton,1951:634-635;Thorndike,1949:第5章)。中间准则比较出名的例子有:(a)大学的分级成绩均值(GPA);(b)是否从大学、培训班等毕业;(c)证书考试成绩。

与终极准则相比,中间准则:(a)更易于界定,(b)更易于测量,(c)获取它们所需要的成本更低,(d)评估它们所需要的时间更短。这些特点使中间准则更具吸引力,但是,如果没有深谋远虑和小心谨慎,我们还是应当避免使用中间准则。选择中间准则时,最重要的考量是它们和所研究的终极准则的相干性。因此,在中间准则上的表现看起来(或应当)和在终极准则上的表现有联系,这常常变成使用中间准则的理据。一般来说,这些主张的基础是逻辑分析,甚或只是预期,而不是经验证据。十分重要的是,我们必须记住:一个既定的中间准则可能和所研究的终极准则没有很大关系或根本没有关系。例如,专科学校(例如,法律、医药)的分级成绩均值和作为专业人员(无论其具体定义是什么)的最终表现之间,很可能没有关系。只要有可能并可行的话,我们就应当收集数据,以便揭示特定中间准则和终极准则之间的关系。

不过,也会出现一些情境,其中,即使我们并不了解中间准则和终极准则之间的关系,中间准则也担当着一个实用的目的。例如,中间准则可以作为有资格或被允许参与某些领域的一个先决条件,这些领域则构成终极准则的各个方面。这样,由于公证和证书的制约,没有从医学院毕业的人,不论他成为“伟大”医生的潜力有多大,社会也会禁止他行医。类似的例子举不胜举。在当前讨论的语境下,是否从专科学校毕业,就是我们希望预测的中间准则。当申请者竞争有限数量的空岗时,当培训在设备、人员、时间等因素上(仅举几个重要因素)投入很大时,这种做法就变得特别

重要,甚至是必需的。

很明显,中间准则的重要性可能会有所变化,如它们的成本、它们与终极准则的关系、它们是不是进入一个既定职业或领域的入场券,都会影响其重要性。因此,我们必须时刻牢记:中间准则的特定角色,仅意味着在特定情境中起作用。这样,我们就可以避免出现下列情形:一个中间准则的"尾巴"摇着一条最终准则的"狗"。

我们已经集中精力,分析了界定和测量准则过程中的内在困难,我们希望借此来强调:准则关联的验证,首先取决于对一个准则的明智选择,然后取决于它的有意义界定和测量。不幸的是,"在准则关联的效度研究报告中,讨论如何评估准则测量本身的报告,几乎凤毛麟角"(Guion,1980:395)。只要这种状况仍然存在,我们就不要指望准则关联的验证研究,能够在应用背景中作出有意义的贡献,更不用说有益于社会行为科学中的理论发展了。

我们在前面曾经指出,准则可能是一个建构,同理,预测因子也可能是。当预测因子、准则或两者都是建构时,在准则关联的验证过程中,有关建构和建构验证(参见第 4 章)的问题和程序,就是关键。这应当是本章开篇所讨论内容的一个提醒,也就是说,分别探讨验证过程的各个方面,虽然比较方便,但我们不能忽视它们之间的相互关系。

## 预　测

在前一部分,我们已经讨论了与准则的定义和测量有关的一些重要的和棘手的问题,这里就不再赘述了。相反,我们假定已经合理地解决了准则问题,而希望利用一个或多个预测因子来预测这个准则。

在这样的情境下,基本的方法是研究预测因子和准则之间的关系。在只有一个预测因子的情形下,预测因子和准则之间关系最常用的指数是"皮尔逊积矩相关系数"。因为我们的讲解仅限于皮尔逊相关系数的使用,因此,把它称为"相关系数"或简称为"相关",就比较方便。在下面的章节中,我们假定大家对相关系数和简单回归分析的基本元素有所了解。我们将在第 17 章中讨论这些主题,当大家对下面的讲解感到困惑不解,或者需要详细推敲时,

请参考第 17 章。

一个预测因子和一个准则之间的相关系数可称为"效度系数"。这样,已知一个预测因子 $X$(例如,智力、焦虑)和一个准则 $Y$(例如,学习成绩、解题能力)的取值,$r_{xy}$ 就是其效度系数。效度系数越大越"好",这句话虽然正确,但我们还是需要记住相关系数的几个要点,下面我们将作简要讨论。

正如第 17 章所讨论的那样,相关系数背后的一个主要假定是,所研究的变量之间的关系是线性的。这意味着,表示个体在两个变量上取值的各个点服从一个趋势,它的特征可以用一条直线来刻画。图 3.1 的(a)便是线性趋势的一个例子,相对照的是,(b)是曲线趋势。举例来说,假定 $X$ = 焦虑,$Y$ = 解题能力。看一下图 3.1 中的两个散点图,我们就会注意到,如果(a)反映情境的话,那么结论就应当是,焦虑增加伴随着解题能力的增加。另一方面,如果(b)反映情境的话,那么结论就应当是,在达到最优点(大约在中等焦虑水平附近)之前,焦虑增加伴随着解题能力的增加;之后,焦虑增加伴随着解题能力的降低。因此,中等焦虑水平似乎最有利解题能力的发挥,而相对高水平的焦虑则会削弱解题能力。

如果数据如图 3.1 中的(b)所示,使用皮尔逊相关系数,就会得到很低的相关系数,由此也会得出一个错误的结论:焦虑和解题能力不相关。很明显,这两个变量是相关的,只不过这种关系不是线性的。底线应当很清楚:在最低限度上,我们也应当用数据作图,看看两个变量之间的关系否有严重偏离线性关系(参见第 17 章)。我们将在第 18 章中"曲线回归分析"的标题下讲解非线性关系的分析方法。

关于相关系数,我们需要牢记的第二个要点是,相关系数随总体而异。在其他条件相同的情况下,研究样本来自的总体越同质,相关系数越低。相反,样本来自的总体越异质,相关系数越高。[①] 我们将在后面的章节再探讨这个问题。

最后,我们只能在比较一般的意义上来阐释一个相关系数。例如,假定两变量之间存在正相关,我们有可能就该效应作出一般性的表述:在一个变量上得分较高的人,有很大的概率在另一个变量上得分也较高;在一个变量上得分较低的人,有很大的概率在另

---

① 我们将用概率抽样来预测这些命题的效度(参见第 15 章)。

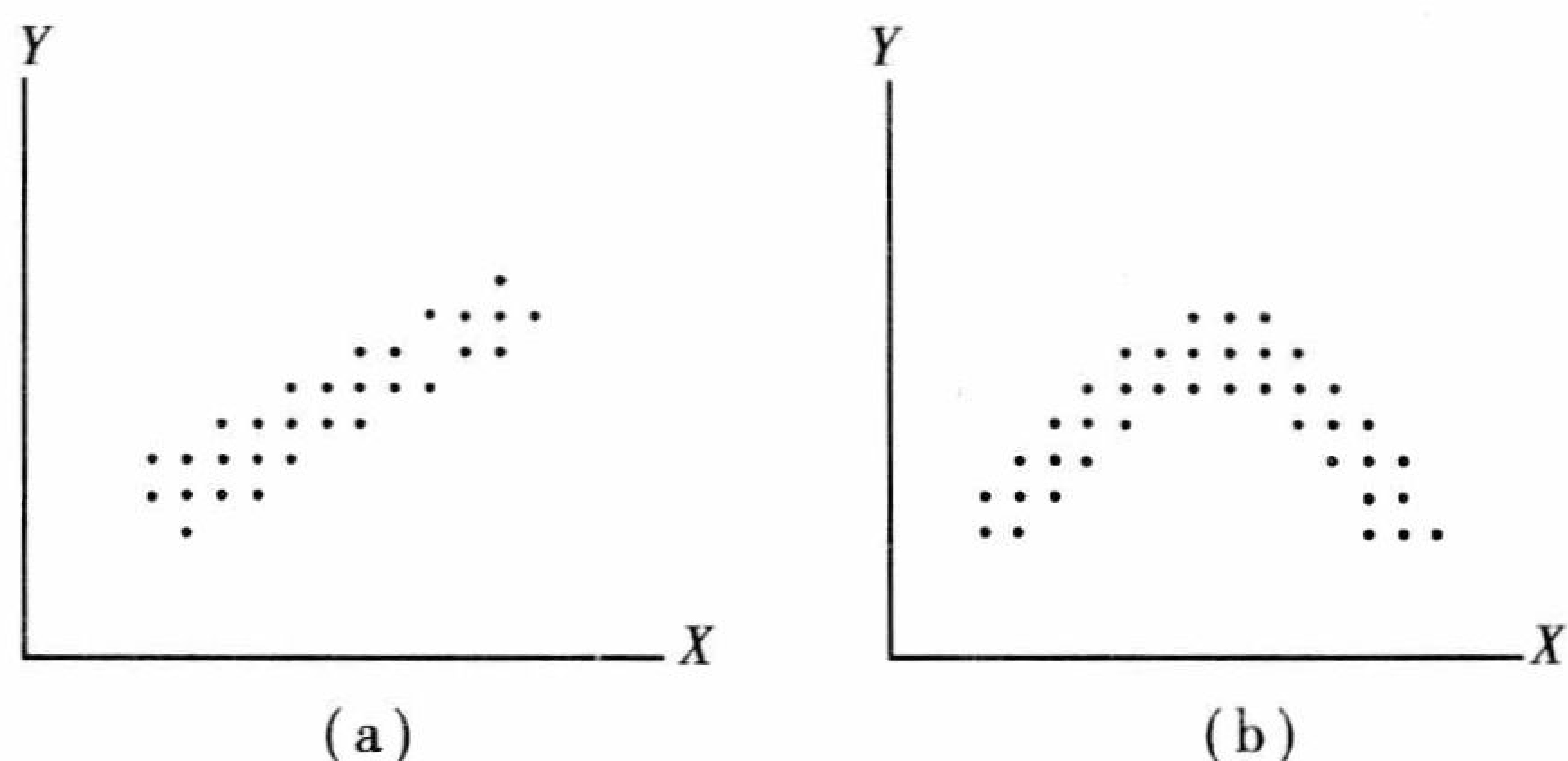



图 3.1

一个变量上得分也较低，等等。然而，就实用性而言（例如，就选择而言），一个预测系统应当让我们能够在预测因子既定的情况下，对准则的预期状况作出更具针对性的预测。实现此目的最有用的方法之一是，使用从回归分析中得到的一个预测方程。

## 预测方程

我们将在第 17 章讨论回归和相关模型之间的区别。就现在的目的而言，我们仅需要指明一点就可以了：只有在回归模型中，我们才区分准则和预测因子（或者，因变量与自变量）[1]，从而得到一个回归方程（其他暂不考虑），让我们可以用一个人在预测因子上的得分，来预测他（或她）在一个准则上的分值。在最简单的情形（例如，只有一个预测因子的线性回归方程）下，这个方程的形式是：

$$Y' = a + bX \qquad (3.1)$$

其中，$Y'$是预测分数；$a$ 是截距，即这条回归线和 $Y$ 轴的交叉点；$b$ 是回归系数或回归线的斜率。

我们将在第 17 章讲解程序，我们可以计算这条预测方程的常数（例如，$a$ 和 $b$），并依据它们，从个体在这个预测因子上的分值来预测他们在这个准则上的得分。此外，我们还有可能计算围绕这些预测分值的置信区间。

在这里，我们的主要关注点是把这条回归方程应用于预测的目的。例如，假定 $Y$ 是生产力的一个指数，$X$ 是一次能力测验中的分数，进行遴选的官员可能会在申请者 $X$ 分数的基础上，使用预测

① 变量的定义和分类，参见第 8 章。

方程来预测他们在 $Y$ 上的分数。而且,一旦他决定了准则($Y$)上的最低限,他也就决定了预测因子($X$)的录取分数线,这样,被录入的人员(即在预测因子上分数高于录取分数线的人)具有令人满意的表现,该事件的概率得到最大化。

这并不暗示:决定预测因子和准则的分数线,是简单、直截了当的事情。这样的决策包含各种考量,我们将在下一节讨论其中的一些考量。这里,我们只想指出:在某些决策中,我们或许可以使用回归方程来进行遴选。

在准则验证的研究中,使用回归方程而不是相关系数,具有一些优势。在讨论它们之前,我们将在"分数线的选择如何影响效度系数的实用性"的背景下,对"预测效率"的概念进行一些评注。

### 预测效率

假定 $Y$ 是我们所研究的一个准则(例如,学习成绩、生产力、康复程度),$X$ 是一个相关的预测因子(例如,能力、人格特质、疾病史)。再假定,不管申请者在 $X$ 上的状况如何,我们都会录取或雇用(例如,上大学或进诊所)他们中的每一个人。一段时间(例如,一年)之后,我们获得了他们在 $Y$ 上的分数。这样,我们就可以计算相关系数和回归统计量。而且,把回归方程应用到 $X$ 分数上,我们就可以获得 $Y$ 的预测分数。

大家可能会问,既然我们已经有了 $Y$ 的实际分数,为什么还要费力劳神,用一条回归方程来预测 $Y$ 的分数?我们这样做,是为了比较实际表现和预测表现,从而确定预测效率。正如第 17、18 章所示,回归分析的各种结果对这种比较都有影响。这里的讲解仅限于介绍一些基本概念,它们和预测效率相关。讲解所借助的图示是图 3.2,其中,点表示 $X$ 和 $Y$ 的分数组合,我们并没有用这些点作图(例如,散点图,参见第 17 章),而是使用了几个椭圆。换句话说,我们假定,所有点都包含在这些椭圆中。

请看图 3.2(a),注意纵坐标轴上的点 $Y_c$。我们用它来表示一个点,高于这个点,我们就认为,在这个准则上的表现是"令人满意的"(例如,成功、康复、毕业)。再看横坐标轴,注意点 $X_c$,它表示在预测因子上的分数线。也就是,如果我们不是录取(雇用)所有申请者,而是利用一条回归方程,那么只有得分高于 $X_c$ 的申请者,才会被录取(雇用)。注意,如果我们从这两个点画两条直线,那么我

们就把这个椭圆分成四个区域，分别表示预测因子和准则状况的四种可能组合。

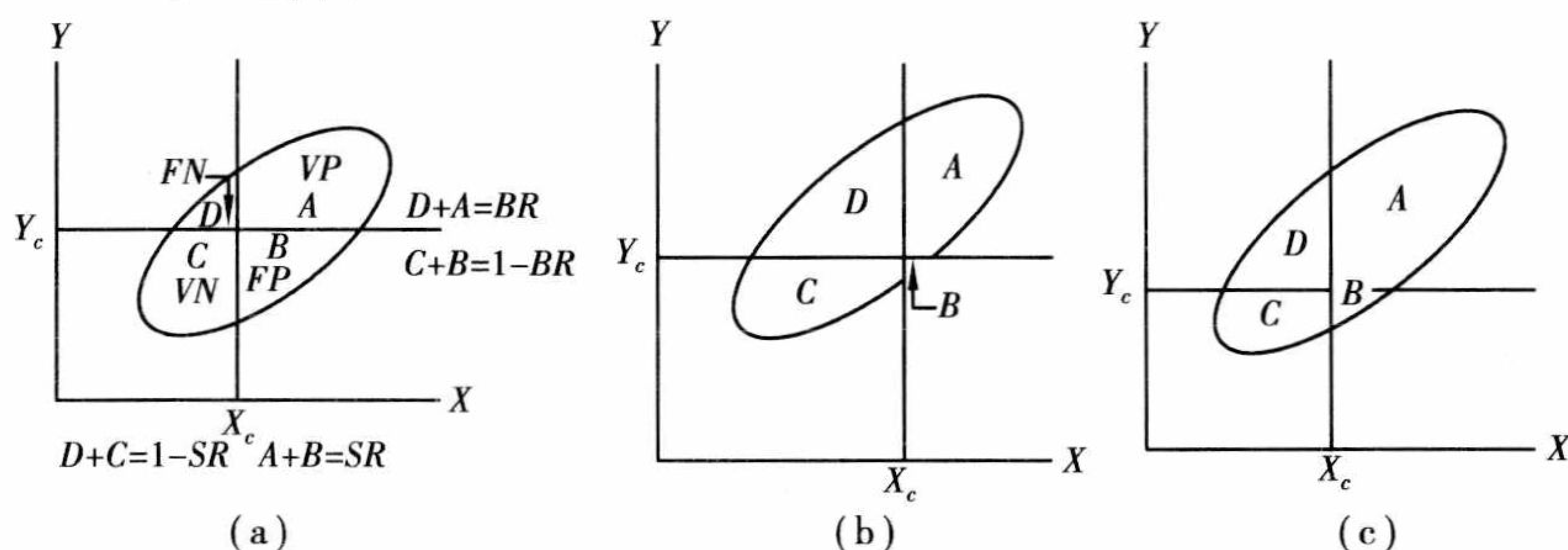



图 3. 2

处在 $A$ 区的人，根据他们在 $X$ 上的得分，我们预测他们能够在 $Y$ 上获得成功（即得分大于 $Y_c$），并且，他们实际上也是成功的；这种状况，我们称之为“真阳性”（$VP$）。处在 $C$ 区的人，我们预测他们不能获得成功，实际上也是不成功的；这种状况，我们称之为“真阴性”（$VN$）。当我们依据一条回归方程进行选择时，这两种情况都可以称为“命中”。

当我们依据一条回归方程进行选择时，其余两个区域都构成“未中”。具体来说，处在 $B$ 区的人，预测的结果是成功，但实际上却是不成功的；这种状况，我们称之为“假阳性”（$FP$）。处在 $D$ 区的人，预测的结果是不成功，但实际上却是成功的；这种状况，我们称之为“假阴性”（$FN$）。

在下面的讲解中，为了方便，我们将采用既定区域或区域组合上的人数比例。例如，当我们表示 $D$ 区和 $A$ 区中的人时，我们将讨论这两个区域中的人数比例，即 $(D+A)/N$，其中，$N$ 表示总人数（其他区域或区域组合，同理）。

这样，当我们录取所有申请者时，$D+A$ 就表示那些表现“令人满意”（即超出 $Y_c$ 以上）的人的比例。这一比例被称做“基率”（$BR$），是指“成功的”人的比例，不管他们在预测因子上的状况如何。换言之，既定的申请者当中，当我们不进行选择，或进行随机选择时，预期可以成功的人的比例就是 $BR$。由于我们使用比例，$C+B$（预期不会成功的人的比例）就等于 $1-BR$。

假定我们只选择分数大于 $X_c$ 的人，那么 $A+B$ 就是入选者的比例，我们可以称之为“入择率”（$SR$）。当然，没有入选者的比例

就等于 1 - *SR*。

参照 *BR* 和 *SR* 来探讨预测效率,其始作俑者是泰勒(H. C. Taylor)和拉塞尔(J. T. Russell),他们把“成功率”定义为 *A* 与 *A* + *B* 之比[即 *A*/(*A* + *B*)],也就是 *VP* 与入选的申请者(即 *VP* + *FP*)之比。(Taylor & Russell,1939)使用上述概念,泰勒和拉塞尔证明,相同的相关系数可以导致较高或较低的预测效率,这取决于 *BR*、*SR* 或两者;而且,“采用较低的相关系数……我们也有可能显著改善选择效率”(Taylor & Russell,1939: 571)。

图 3.2 是泰勒和拉塞尔所提出的基本观点的图示,其中,三个椭圆描述的是相同的散点图。椭圆的宽度表示预测因子和准则之间的相关系数(例如,效度系数)相对较低。(粗略地说,椭圆越窄,相关系数越高。当所有的点都落到一条直线上时,相关系数当然就是 1,最完美的情形。参见第 17 章。)

让我们首先考察:在其他条件保持不变的情况下,降低入选率(*SR*)所产生的效应,对比图 3.2 的(a)和(b),我们便可以看到这种效应的一个示例。其中,图 3.2(b)中的 *A* + *B* 小于图 3.2(a)的 *A* + *B*。注意,在图 3.2(b)中,我们采用了较高的预测因子分数线,才得到这种效应。这样,和图 3.2(a)相比,图 3.2(b)中的假阳性(*FP*、*B* 区)的比例较低。按照泰勒和拉塞尔对“成功率”的定义(参见前文),很明显,(b)中的成功率大于(a)。

沿着 *X* 轴向右移动分数线,即采用更小的 *SR*,我们有可能完全消除假阳性,达到 100% 的成功率(即成功率等于 1.00)。这样,相对于一个既定情形下的申请者总数而言,能够入选的申请者数量会变得很少。而且,*SR* 的降低也会影响到 *FN* 和 *VN* 的比例。特别需要注意的是,*SR* 的降低会导致 *FN* 比例的上升——我们将在下面讨论这个问题。

现在让我们比较图 3.2(a)和(c)。注意,在这两张图中,*SR* 相同,但图 3.2(c)中的 *BR* 较大。$Y(Y_c)$ 是界定成功的相对次要的因素,降低 $Y(Y_c)$ 的分数线,我们就可以得到较大的 *BR*。如前所述,*BR* 是成功者的比例,不过,他们在预测因子上的状况并没有用来进行选择。请注意,*BR* 越大(即合格的申请者比例越大或 *Y* 上的分数线越低),*FP*(*B* 区)的比例就越小;不过,*FN* 的比例也会有一定的增加。

同时还应注意,*BR* 越大,预测因子的效应就变得越小。在极

端的情形下，当$BR=1.00$（即所有申请者都合格，或者，无论多低，任何表现都被认为是令人满意的）时，预测因子毫无用处。在这类情况下，如果还有必要进行选择的话（例如，当职位空缺数小于申请者数时），那么随机选择很可能是最公正的办法。

假定 $X$ 分数和 $Y$ 分数是双变量正态分布，泰勒和拉塞尔研发了一些交互表，其中，成功率（参见前文）是变化的效度系数（$r_{xy}$）、$BR$ 和 $SR$ 的函数。（Taylor & Russell，1939）表 3.1 是几个例子，摘自泰勒和拉塞尔的表格，其中，在（a）列中，$SR$ 和 $BR$ 保持不变，$r_{xy}$ 发生变化；在（b）列中，$SR$ 和 $r_{xy}$ 保持不变，$BR$ 发生变化；在（c）列中，$BR$ 和 $r_{xy}$ 保持不变，$SR$ 发生变化。

**表 3.1　泰勒和拉塞尔表格的摘录**

| (a) | | (b) | | (c) | |
|---|---|---|---|---|---|
| $SR=0.50$ | | $SR=0.50$ | | $BR=0.50$ | |
| $BR=0.50$ | | $r_{xy}=0.40$ | | $r_{xy}=0.50$ | |
| $r_{xy}$ | | $BR$ | | $SR$ | |
| 0.20 | 0.56 | 0.10 | 0.16 | 0.10 | 0.78 |
| 0.30 | 0.60 | 0.30 | 0.41 | 0.30 | 0.69 |
| 0.60 | 0.70 | 0.70 | 0.81 | 0.70 | 0.58 |

注释：$SR$ = 入选率；$BR$ = 基率；$r_{xy}$ = 效度系数。表格中的数字是成功率。具体解释，参见正文。

表格中的数字是成功率。例如，当 $SR=BR=0.50$，且 $r_{xy}=0.20$ 时，查看表格（a）列的第一行，我们就可以得到成功率是 0.56。这意味着：使用预测因子时，成功率的预期增幅是 0.06 或 6%（在当前这个例子中，如果不使用预测因子的话，预期的成功率应等于 $BR$，即 0.50）。可见，即使是一个相对较低的效度系数，也能让成功率上升，在各种情境中，这个上升幅度也会被大家认为是有意义的。

现在，让我们来比较（a）列和（b）列的第一行。在这两种情形下，成功率的增幅都是 0.06（如果不使用预测因子的话，对 $BR=0.10$来说，预期的成功率是 0.10）。不过，在（a）列中，效度系数是 0.20；而在（b）列，效度系数却是 0.40。这演示了上文所说的内容，即一个既定效度系数的预期绩效依赖于其他因素（即 $BR$ 和 $SR$）。在其他条件相同的情况下，当 $BR=SR=0.50$ 时，成功率的增幅达到最大。

当 $SR = 0.50$, $r_{xy} = 0.40$, $BR = 0.30$(参见(b)列)时,成功率是 0.41,和不是基于这个预测因子的选择过程相比,其增幅是 0.11 或 11%。请注意,当 $BR$ 和 $r_{xy}$ 保持不变时(参见(c)列),$SR$ 越低,成功率越高。例如,当 $SR = 0.70$ 时,成功率的估值是 0.58;当 $SR = 0.10$ 时,成功率的估值是 0.78。[①]

所谓"成功"是指,在这个选择过程中,由于使用预测因子而带来的 *VP* 比例的增加。这是泰勒和拉塞尔从预测效率的视角所定义的成功,它也应用于前面的各种示例中。一些雇主或机构对 *VP* 最大化感兴趣,在他们的眼中,这种定义或许是最有用的,但是,它的应用也需要付出成本。

例如,伯克森(Joseph Berkson)将"成本"定义为一种比例,有些申请者应当是成功者,但由于他们在预测因子上的得分低于分数线,因而被预测为不成功者(也即假阴性;参见图 3.2 的 *D* 区),这些人占申请者的比例就是"成本"(Berkson,1947)。先前我们曾指出,降低 *SR*(入选率)会导致成功率的增加,同时,*FN*(假阴性)的比例也会随之升高。依据申请者在预测因子上的得分,有些人会被拒绝;但如果雇用或录取他们的话,这些人将可能成为成功者;对他们而言,不必说,*FN* 尤其受到关注。他们肯定会挑战泰勒和拉塞尔对成功的定义,同理,有些人也会关注:如果不能人尽其才,社会就会付出成本。如前所述,在其他条件保持不变的前提下,降低 *SR* 将导致 *VN*(真阴性)的增加,伯克森(Berkson,1947)将 *VN* 称为"效用"。

随着一个既定研究的特定目的的不同,有时候,我们可能会直接投入精力,最大限度地减少某些类型的错误。例如,洛伯(Rolf Loeber)和迪肖恩(Thomas J. Dishion)采用上述方法,综述了预测青少年犯罪的各种研究。他们指出,就有罪的司法裁定而言,我们应当将假阴性降到最低:"也就是说,预测因子不应当放过那些实际上犯罪的青少年。"(Loeber & Dishion,1983:70)另一方面,如果我们关注校正或预防,那么我们就应当把假阳性(即那些"看似具有

---

① 吉塞利等人沿用了泰勒和拉塞尔的表格(Ghiselli et al.,1981),对这里所讲解的概念,他们进行了很好的讨论。关于这些概念的其他讨论,参见艾伦和耶尼(Allen & Yen,1979:101-107)、威金斯(Wiggins,1973:240-250)。内勒和希恩(Naylor & Shine,1965)拓展了这里所讨论的概念,并提供了一些表格,供我们利用预测因子来查找准则的均值(取代泰勒和拉塞尔表格中的成功率)增幅。

犯罪风险,但尚未犯罪的青少年")降到最低。(Loeber & Dishion, 1983: 70)

上述讲解旨在对预测和预测后果的定义等复杂问题作一个介绍。我们的主要目的是举例说明,完全依赖效度系数是不明智的;有关预测效率和成功的各种判断依赖于这些概念的特定定义。更多的讨论和扩展,参见克伦巴赫(Cronbach,1971)、米尔和罗森(Meehl & Rosen,1955)和西克里斯特(Sechrest,1963)。克伦巴赫和格莱泽(Cronbach & Gleser,1965)对测验在选择中所起的作用,进行了广泛和细致的讨论。

## 范围限制

为了详述一个选择过程的各种可能后果,在前面的章节中,我们假定,所有申请者都将得到录取或雇用。由于一些很明显的原因,这样的研究几乎无法进行。在大多数情形下,总会出现一些选择,并且,效度系数是利用选择组的分数计算得出的。一般而言,由于选择组比全体申请者更同质,利用从前者所获得的数据、计算出的效度系数,比利用从后者所获得的数据、计算出效度系数,会要小一些。

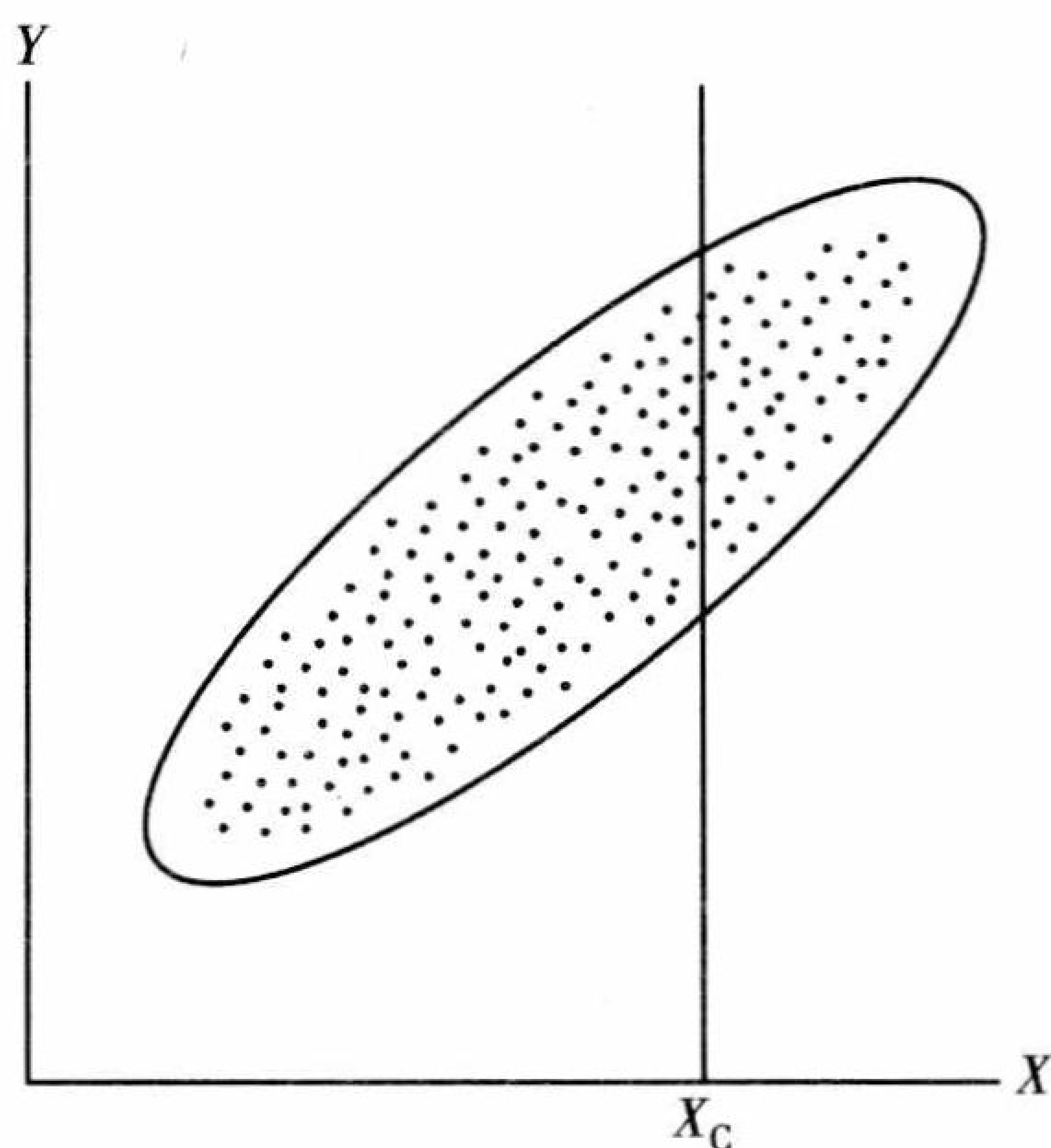



图 3.3

(具体示例,参见 Linn & Dunbar,1982)图 3.3 是这种情形的一个图示说明,其中,$X$ 代表预测因子,$Y$ 表示准则。当我们能够得到所有申请者的数据时,整个散点图就是对预测因子和准则之间关系的描述。但是,当我们只能得到预测因子的分数线($X_c$)之上的申请者的数据时,那么这个散点图就类似于图 3.3 中竖线右边的区域。基于这个区域内数据所得到的相关系数,会小于基于整个散点图的数据所得到的相关系数。由于选择而造成一个相关系数缩小其幅度,一般被称做"范围限制"对这个相关系数的效应。

总的来讲,存在三种范围限制:直接、间接和模糊。当我们在预测因子的分数线基础之上进行选择时,就会出现"直接限制"。例如,一所大学只招收那些在"学术能力测验"的语文部分(SAT-

V)得分超过分数线的学生。这是最简单的一种范围限制形式(参见图 3.3 的图示)。现在,设想一种情形:一所大学招收了一批学生,他们在高中时的分级成绩均值(GPA)高于分数线;后来,校方想要利用 SAT-V 成绩来预测这批入学的学生在大学的分级成绩均值(GPA)。因为高中的分级成绩均值和学术能力测验成绩倾向于正相关,所以,对前者的范围限制也倾向于限制后者的范围,这就是"间接限制"所表示的过程。

不过,范围限制的最常见情形是前文所说的"模糊限制"。这意味着,尽管我们可能对运作在一个选择过程中的变量有怀疑或一些了解,但是,对这些变量以及它们如何组合起来、产生选择组的过程,我们却不拥有充足、有效的信息。例如,一所大学可能会采用 SAT-V 的分数线作为入学条件,但却没有严格执行这条分数线,出现了很多例外,例如说,考虑少数民族身份、课外活动、与大学校友的关系等。

一种更为棘手,且经常出现的情形是"自选择"过程出现的时候。例如,一所大学可能会严格执行 SAT-V 分数线,但在得到通知、告之他们的入学申请已经得到批准的申请者当中,只有一部分人会选择入学。

尽管存在用于校正范围限制的公式(例如,参见 Ghiselli et al, 1981: 296-306; Lord & Novick, 1968: 140-148; Thorndike, 1982: 208-215),但是,它们只适用于非常简单、有限的情境。例如,林(Robert L. Linn, 1968, 1983a, 1983b)曾就"范围限制"这个主题作过广泛的讨论,他指出,在前文称做"模糊限制"的条件下,校正相关系数的尝试,可能会带来严重的偏差。

最后,关注回归方程比关注效度系数更有意义。我们将在下一章详细讨论这个问题。现在,我们仅指出,在某些条件下(即直接限制),由于范围限制,效度系数将发生变化,而回归方程却会保持相对稳定。让我们回过头来看一下图 3.3,目测或动手画一条 $Y$ 对 $X$ 的回归线,请注意,无论是在整个散点图上,还是在分数线之上的区域,这条回归线大体相同。实际上,对范围的直接限制而言,校正其相关系数的公式的基础是一个假定:当预测因子的取值范围受到限制时,回归系数($b_{yx}$)保持不变(具体示例,参见 Ghiselli et al, 1981: 296)。

## 分组预测

在前面的章节中,我们对预测效率的讨论仅限于一种情形,即所有申请者都被当做是隶属于同一个组别。不过,有时因为要求,有时出于兴趣,我们也会依据申请者在不同既定组别中的成员隶属,区分不同的申请者。对于准则关联的验证而言,我们有时需要确定:对不同的组别(例如,男性和女性,或者白人、黑人和西班牙裔),一个预测因子是否具有相同的预测力。在选择过程中,针对一定既定组别的成员,是否存在偏见(参见下文)? 如果这是我们的目的,那么,区分申请者就变得至关重要。可见,当我们把申请者看做是隶属于多个组别时,我们所关注的就是分组预测所关注的。

有大量的文献探讨了所谓的"单组效度"和"多组效度"。因为前一个概念存在不当表达(Cronbach,1980),我们就不做讨论了。"多组效度"是指我们从不同组别得到的效度系数之间的差异。在这里,我们并不引用有关多组效度的文献,因为多组效度仅仅是不同组别之间的可能差异的一个方面,这些差异都可能会导致分组预测。在下列由 $X$ 预测 $Y$ 的回归方程的表达式中,我们可以看出这一点:

$$Y' = \left[\bar{Y} - r_{xy}\frac{S_y}{S_x}\bar{X}\right] + \left[r_{xy}\frac{S_y}{S_x}\right]X = a + bX \tag{3.2}$$

其中,$Y'$是 $Y$ 的预测值,$\bar{Y}$ 和 $\bar{X}$ 分别是 $Y$ 和 $X$ 的均值,$r_{xy}$是效度系数,$S_y$ 和 $S_x$ 分别是 $Y$ 和 $X$ 的标准差。如公式 3.2 所示,第一项(中括号中)表示截距($a$),第二项(中括号中)表示回归系数($b$)。[①] 由公式 3.2 可见,在预测因子($X$)的基础上预测准则($Y$)的分值,效度系数($r_{xy}$)只是这个回归方程中的元素之一。因此,单纯依赖 $r_{xy}$,虽然在多组效度的研究中常见,却不是明智之举。

分组预测是指不同组别的回归方程之间的差异。不同组别可能具有不同的回归系数($b$)和(或)截距($a$)。由公式 3.2 可见,对两个或以上的组别来说,效度系数相同,但回归系数可能会有所不同,差异的大小取决于标准差。反之,不同组别的($b$)相同,但效度

---

① 对回归方程的计算和阐释的详细讨论,参见第 17 章。

系数可能会有所不同。因为截距($a$)是($b$)和均值的函数,由此推论,在回归方程中具有相同截距的不同组别,可能具有不同的回归系数和(或)效度系数。

为了厘清上述观点,我们将在图 3.4 中举一个相对简单的示例。假定这个图表示女性和男性在预测因子($X$)和准则($Y$)上的分数。仔细看看这两个点团,目测两条拟合它们的回归线(大家可能会发现,动手画出这两条线,也十分有用)。显然,两组的回归系数($b$)非常相似,截距($a$)相差很大。而且,女性的效度系数比男性的要小一些(女性比男性的点团要宽一些)。

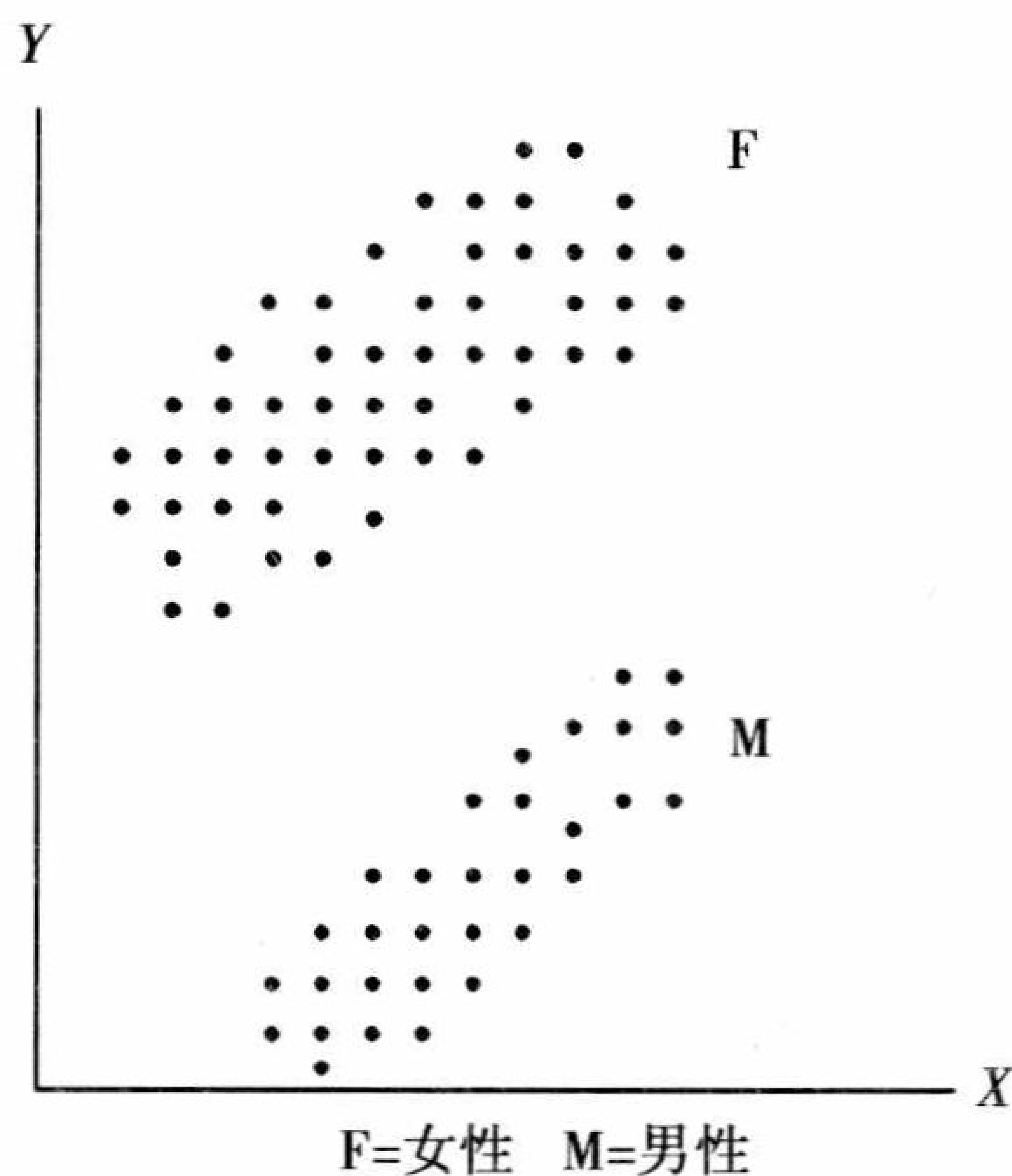



图 3.4

现在,只看效度系数的话,我们得出的结论是:对于男性来说,$X$ 是更有效的预测因子;但如果用各自的回归方程的话,预测的结果是女性的表现要高于其男性对手。假定结论是两者的($b$)相同,那么,对任意既定 $X$ 而言,女性的 $Y$ 预测值将高于男性的 $Y$ 预测值,差异的幅度等于这两个截距间的差异(参见下一节"比较回归方程")。

图 3.4 中的示例也说明,当被研究的对象属于不同的组别,以某些相关的方式,他们在所研究的变量上存在差异时,采用单个相关系数或者采用一个回归方程,都会带来潜在的风险(参见下面的"研究分组的性质"和"选择偏差")。

## 比较回归方程

在前一节中,我们用目测的方式比较了两个组别的数据。我们将在第 21 章讲解回归方程间的差异检验。就当前的目的而言,我们仅就如何进行检验作一个简短的说明。① 我们的起点是检验 $b$ 间的差异是否具有统计显著性。请注意,$b$ 表示回归线的斜率,由此推知,得出"$b$ 间差异统计上不显著"的结论,相当于得出"两条回归线平行"的结论。在这样的情形下,继续进行第二阶段的分析(即截距间的差异检验)才是有意义的。如果我们发现截距间的差异具有统计显著性,那么,我们就要分别拟合不同的回归方程,对所有组别而言,$b$ 相同(即一个共同的或组间合并的 $b$;参见第 21 章),$a$ 却彼此不同。另一方面,如果第二阶段分析所得出的结论是"$a$ 间差异统计上不显著",那么,对所研究组别的所有成员,拟合和采纳一个回归方程,就可以了。

正如第 21 章所探讨的,当结论是"$b$ 间差异具有统计显著性"时,截距间的差异检验也就没有意义了。取而代之,我们应给涉及的每个组别分别拟合一个回归方程。当我们需要预测一个既定个体的准则分数时,我们就使用他(或她)所隶属的那个组别的回归方程。

我们将采用图 3.5 来进一步说明这些观点。为了避免画面散乱,我们并没有在图中显示表示个体分数的点,而只画了回归线。为了便于讲解,我们只使用了两个组别和一个预测因子。正如在第 21 章中所讨论的,这里所描述的方法,也适用于任意数量的组别和(或)任意数量的预测因子。

图 3.5 显示了回归方程间可能存在的各种差异。A 和 B 两个标签表示任意两个组别(例如,男性和女性,黑人和白人,律师和警官)。在每种情形下,这些回归线分别反映了两个组别各自的回归方程。为了方便讨论,我们在图中标明了在预测因子($X$)上具有相同分数个体的预测分数($Y'_A$ 和 $Y'_B$)。采用各自的回归线来表示既定 $X$ 值的预测分数,相当于从 $X$ 值处画一条垂直线,与 A,B 两条回归线相交,然后得到 $Y'_A$ 和 $Y'_B$。

---

① 依据个人的知识背景,大家可能会发现,对第 21 章相关部分的阅读有助于更好地理解这一部分。

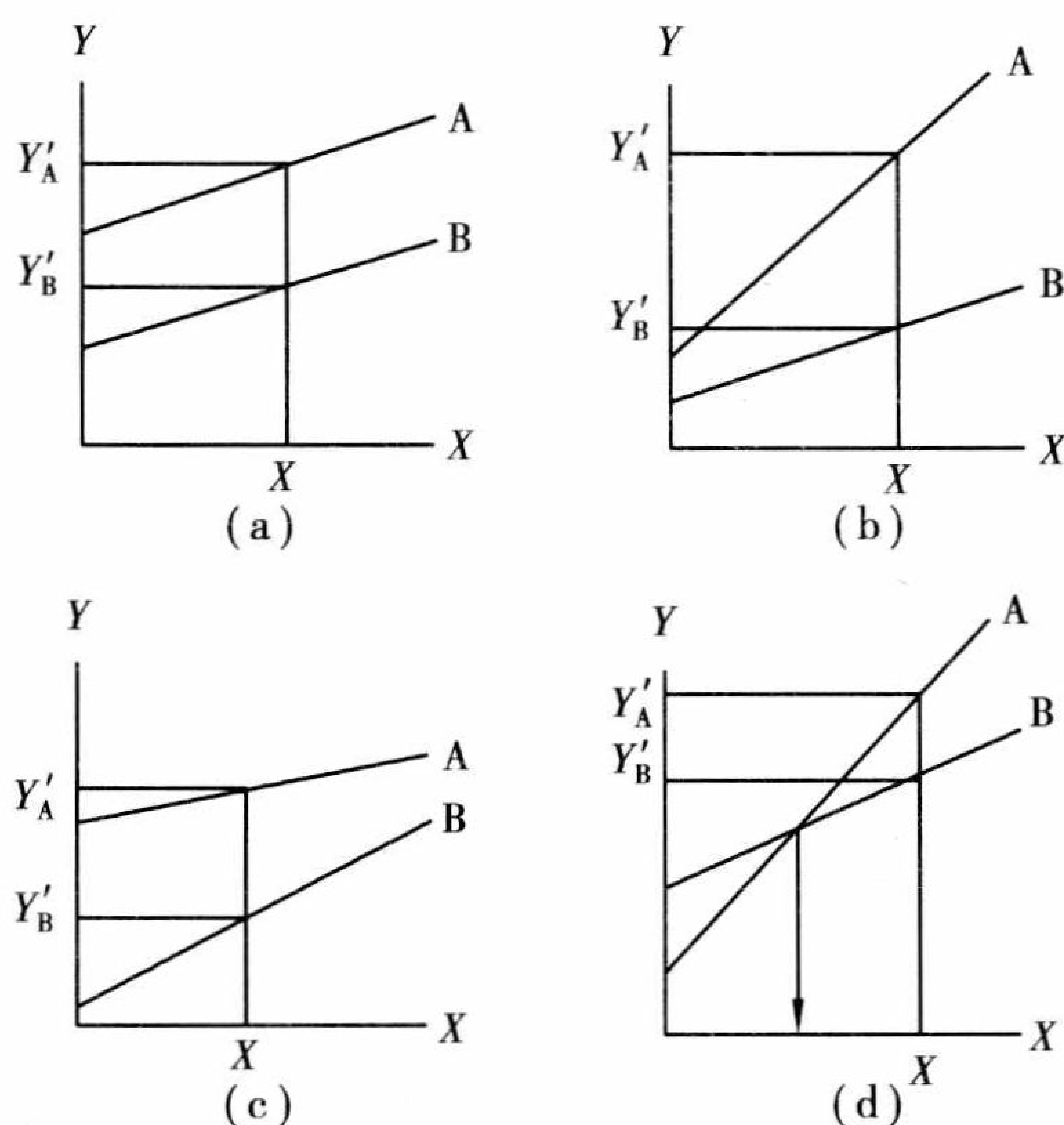



图 3. 5

首先,让我们看一下图 3. 5(a),请注意,两条回归线平行,这表明:两个回归方程具有相同的回归系数(b)。但两个截距(a)间的差距相对较大。在这种情况下,A 组成员的预测分数总会比 B 组成员的预测分数大一些。对任意既定的 $X$ 值,预测分数间的差异就等于这两个截距间的差异。图 3. 4 所表示的,也是这个类型的例子,它也表明,两个组别的效度系数有差异。

现在转到图 3. 5(b)和(c),请注意,在这两种情形下,A 的回归线高于 B 的回归线。当然,这意味着,对于任意既定的 $X$ 值,A 组成员的预测分数比 B 组成员的预测分数要大一些。但预测分数间的差异幅度,取决于 $X$ 的特定取值。在图 3. 5(b)中,预测分数间的差异会随着 $X$ 值的增加而逐步增大。在图 3. 5(c)中,情形恰恰相反。再强调一次,$A$ 的 $r_{xy}$ 与 $B$ 的 $r_{xy}$ 可能相同,也可能不同。

用"属性—处理—互作"(Attibutes-Treatment-Interaction, ATI)设计的术语(参见第 12 章和第 21 章)来说,图 3. 5(b)和(c)所表示的情形,可称做预测因子和组别隶属间的"同序互作"。简而言之,这意味着,尽管一个组别成员的预测分数总是大于另一个组别成员的预测分数,但差异幅度取决于预测因子($X$)的特定取值。对比图 3. 5(b)、(c)和(a):在图 3. 5(a)中,预测因子和组别隶属之间

并不存在互作，因此，在图 3.5(a)中，预测分数间的差异是一个常数；而在(b)和(c)中，它们会随着 $X$ 取值的变化而不断变化。

最后，让我们转到图 3.5(d)，请注意，两条回归线交叉——这种情况被称做预测因子和组别隶属间的“异序互作”(参见第 21 章)。它所说明的是，一个组别的预测分数并不总是大于另一个组别的预测分数，这和我们所描述过的各种情形都不同。例如，在图 3.5(d)中，对应于两条回归线交点(箭头所示)的 $X$ 值，当人们在预测因子上的取值低于这个值时，$B$ 的预测分数高于 $A$；当人们在预测因子上的取值高于这个交叉点时，则相反的情形成立。

在讨论图 3.5 中所示的情形时，我们把注意力放在回归方程上。如前所述，就回归方程间的特定比较而言，效度系数($r_{xy}$)可能相等，也可能不相等。这并不是说，相关系数的大小无关紧要。首先，如前所述，$r_{xy}$ 是决定 $b$ 的幅度的因素之一。其次，在决定估值的标准误($s_{y.x}$，参见第 17 章)时，因而在决定预测分数的置信区间时，相关系数也起一定作用。(有关讨论、示例和参考文献，参见 Pedhazur，1982：143-147)

就当前的目的而言，我们只需要指明，在其他条件相同的前提下，相关系数越小，$s_{y.x}$ 就越大，预测分数的置信区间因而就越宽。大概来说，一个预测分数的置信区间越宽，人们对基于预测分数所作出的决策的信心就越小。(关于这个论点的较好讨论，参见 Einhorn & Bass，1971；同时参阅 Barrett，1974)

## 研究分组的性质

我们可以把人们分为各种各样的组别，这样，一个问题便出现了：为了比较回归方程，我们如何决定采用什么样的分组或分类变量？对这个问题，简单的答案并不存在，我们只能说，这取决于理论命题和(或)特定兴趣。例如，对基于一个预测因子的一个选择程序，如果我们怀疑它对女性有偏差，那么，我们就需要比较男性和女性的回归方程。

## 选择偏差

在预测因子分数基础上遴选申请者时，如果有人提出异议，认为这种选择偏爱或歧视一个组别的成员(例如，女性，黑人)，那么，我们就要对“选择是否有偏”进行论辩。由于立法、司法和大众的

关注(有关综述,参见 Bersoff,1981),“选择偏差”的问题已经得到了社会行为学家,特别是测量专家的不断增长的注意。正如其他研究领域一样,关于“选择偏差”,也不存在一个共同的、广泛接纳的术语(更不用说术语的定义了)。例如,一些研究者交替使用诸如“选择偏差”“测试偏差”“测试公正”等术语,其他研究者则对它们进行了清晰的区分。(参见 Petersen,1980)

我们的目的并不是回顾有关选择偏差的大量文献和相关问题,我们也不想探讨与这个主题有关的各种定义、测量和统计问题。(例如,参见 Arvey & Faley,1988;Berk,1982; Cole,1981;Cole & Moss, 1989; Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, 1978; Green, 1981a, 1981b; *Journal of Educational Measurement*, 1976,13:春季卷; Pezzullo & Brittingham,1979; Reynolds, 1982; Reynolds & Brown,1984)在这里,我们仅指明,选择偏差的一些处理不仅在技术上很复杂,并且都是建立在外显或内隐的价值判断(即“歧视”“公正”的意义)之上。(可参阅 Flaugher,1978; Hunter & Schmidt,1976; Petersen & Novick,1976)除了提醒大家,选择偏差是一个复杂的论题之外,我们的目的仅限于说明,研究选择偏差的最常见方法之一,它的基础是我们在“分组预测”这一节中所讲解的观点。

克利里(T. Anne Cleary)提出了下面的定义:

> 为了预测一个准则,我们设计了一个测试。对总体中的一个子群体的成员而言,如果预测这个准则时,该组成员出现一致的非零误差,那么,对他们来说,这个测试就是有偏的。换言之,从一个公共的回归线所预测的准则值,对这个子群体的成员来说,如果总是一致性地过高或过低的话,那么,这个测试就是有偏的。(Cleary,1968: 115)

克利里的定义被称做“回归模型”①,这并不会让我们感到惊奇,因为这个定义建立在比较上述回归方程的基础之上。为了清楚说明这一点,我们再来看一下图 3.5。首先,让我们看一下图 3.5

① 克利里所说的“测试偏差”,许多学者也称做“选择偏差”。在前面,我们已经提到缺乏通用术语的问题。

(a),请注意,公共的回归线就是和 A,B 两条回归线等距、平行的一条直线(大家会发现,动手画出这条线,会非常有用)。使用这条公共的回归线(而不是分别使用两条回归线)来进行选择时,A 组成员被一致低估,B 组成员被一致高估,这样,就出现了克利里所定义的“偏差”。

在图 3.5 所表示的其他情形中,使用一条公共的回归线,也会导致选择偏差的出现,但情况会更加复杂。在图 3.5(b)和(c)中,一条公共回归线会导致对 A 值的低估,对 B 值的高估,但偏差的程度取决于 $X$ 的取值。在图 3.5(b)中,$X$ 越大,偏差越大。在图 3.5(c)中,情况恰恰相反。最后,在类似图 3.5(d)所显示的情形中,当 $X$ 的取值低于两条回归线的交点时,使用一个公共回归线,可能会导致一种有利于 B 组成员、不利于 A 组成员的偏差。对于在 $X$ 上的分数高于交点的人来说,相反的情形成立。

### 多个预测因子

本章的讲解局限于单个预测因子的情形。许多情况下,我们用到一个以上的预测因子。此时,多重回归分析(参见第 18 章)是最常用的分析方法。而且,选择预测因子的各种方法(即逐步回归分析)也有了长足发展。(有关综述和参考文献,参见 Pedhazur, 1982:第 6 章)

## 结束语

我们相信,重述本章开篇时所说的话是非常重要的,即为了方便起见,我们在不同的章节中讲解了准则关联的验证和建构验证,但它们构成了一个验证过程的不同维度。而且,可以确定的是,在本章简要提及的各种问题,特别是那些涉及分析方法的问题,都会在随后的章节中得到详细讨论。因此,我们简要提到的一些主题,如果有些模糊的话,也不必太在意。如果此时您想厘清我们所提到的某个特定主题的话,建议您阅读第 3 部分(特别是第 17、18 和 21 章中)的相关论述。

# 4

# 建构验证

在前一章我们集中探讨了利用一个或多个预测因子的信息来预测一个准则的各种问题。如前所述,在这些探讨中,尽管理论考量绝不是毫不相干,但它们也没有扮演一个中心角色。当我们转到建构验证时,这种观点就要彻底改变一下了,因为一个建构的定义和意义都来源于它所根植的理论网络。

让我们从考察利用指标来对建构进行推论开始。然后,在逻辑分析、内结构和跨结构分析、趋同验证和判别验证的标题下,介绍建构验证的方法。本章以对内容效度的评论结束。

## 建构和指标

建构是"概念"的同义词,是理论建构、抽象,旨在组织和解释我们的环境。换句话说,建构"既不是一种视觉形象,也不在大脑之外;它像是思想所把玩的一种游戏中的一个道具"(Caws,1959:16)。焦虑、动机、智力、态度、自尊、兴趣、挫折和利他等,都是建构的示例。

"建构验证"是指在可观察变量(假定是建构的指标)的基础上对不可观察变量(建构)进行推论的效度检验。康德这样表述建构和指标之间的互惠关系:"没有事实内容的概念是空洞的,没有概念的感觉资料是盲目的。"(转引自 Mackay,1977:84)建构验证隐含着各种困难、模棱两可,甚至是循环论证。人们很可能会问:一个既定的可观察变量是一个本身不可观察的变量的指标,我们如何才能确定这一点?而且,一个既定的可观察变量可能反映不同

的建构(即相同的行为可能反映不同的动机),相同的建构可能显现在不同的可观察变量之上(即相同动机可能由不同的行为所反映);考虑到这个事实,我们如何才能区分它们?

下面这则新闻,说明了从一个指标向一个建构进行推论的过程中所面临的内在模糊性:

> 我们听说,某个东部城市的博物馆以其惊人的参观人数引以为自豪。近来在这个博物馆的附近建造了一个石头小建筑。次年,到这个博物馆的参观人数,令人奇怪地下降了10万人次。这个石头小建筑是什么呢?一个舒适的车站而已。(*This Week*,1948年4月17日,转引自Wallis & Roberts,1956:133)

下列事实让事物变得更复杂:在某些既定情境下,一个可观察变量本身是研究兴趣所在;在另一些情境下,它又被看成是某些建构的一个指标。因此,举例来说,我们可以研究投票行为自身,也可以把它看成是一些建构(例如,政治参与)的一个指标。或者,在一个既定的研究中,研究兴趣可能是受教育程度对收入的效应;而在另一项研究中,这两个变量都可能被看成是社会经济地位的指标。

当我们把可观察变量作为一个建构的指标时,我们一定要十分谨慎,不能把归属于这个建构的意义,充塞到这些变量之上,否则的话,我们就会得出错误的,甚至荒谬的结论。例如,《教育机会的公平性》(Coleman et al,1966,通常也被称作《科尔曼报告》)是一个有影响力的研究,在对它的数据进行再分析之后,阿莫尔(David J. Armor)发现,由9件家庭物品(即电视机、真空吸尘器、电话、辞典、冰箱等)的拥有率所构成的一个指数,和学生的语文成绩之间存在大约0.7的相关系数(Armor,1972)。现在,将上面提到的物品拥有率看成是一个建构(例如,"家庭生活方式""一个家庭的经济福祉",参见Armor,1972:206)的指标,并对这个相关系数作相应的阐释,这是有意义的。我们假定这些指标表达了这个建构,但如果我们把这个建构的意义带到这些指标的身上,很明显,这是完全不同的两码事,因为这可能会导致这样一个结论,即拥有电话、冰箱、真空吸尘器等会影响学生的语文成绩。

这个例子过于明显,好像不值一提。但类似这样的错误概念

却普遍存在。例如,在报告“教育成就的国际研究”(有关 IEA 研究,可参阅 Peaker,1975)的发现时,赫钦格(F. M. Hechinger)认为:“与学生的家庭收入和受教育程度相比,学生家中的书籍和杂志数量对文学成就的影响要大得多。”(Hechinger,1973)依此来看,不用细推,我们就会注意到,家中的书籍和杂志的数量也公认是家长的受教育程度或收入的指标;更不用说,前面提到的所有物品,也可能是社会经济地位的指标。

### 时间、地点和情境的考量

由于本身的特性,随着地点、文化、亚文化等的不同,一个指标可能具有不同的意义。此外,在一个既定地点,由于历史事件,规范、经济条件(仅举数例)等变迁,指标的意义也可能在时间进程中发生变化。当指标是对诸如态度或人格问卷中的题器所作的应答时,尤其如此。作为一个示例,让我们转向社会心理学中最有影响力的一项研究,即《权威人格》(Adorno et al,1950),它的一个主要方面是开发“F 量表”(法西斯主义量表),这是后来广为人知的“权威主义量表”。在这里,讨论和研究《权威人格》的作者所使用的“权威主义”这个建构,既不可能,也无必要。(相关的综述及批评,参见 Christie & Jahoda, 1954; Kirscht & Dillehay, 1967; Sanford, 1973)

就当前的目的而言,F 量表中的题器具有意义上的差异,我们只要注意到它们的一些来源就行了。在 F 量表的早期研究之中,克里斯蒂(Richard Christie)和加西亚(John Garcia)发现,对不同地区的美国大学生而言,F 量表的题器具有不同的意义,有些差异可以归因为亚文化差异(Christie & Garcia,1951)。最近,米勒(Miller et al,1981)等人指出,在美国和波兰,F 量表的一些题器似乎触及相同的维度,其余题器则更加文化有别。

F 量表的题器意义的最明显变化,源自美国社会中的重大事件(例如,水门事件、伊朗军售案、妇女解放运动、同性恋解放运动)。由于这些事件、意义发生明确变化的 F 量表题器如下:

同性恋者同犯人一样,应该受到严惩。

大多数人并没有意识到我们的生活在很大程度上受控于暗处密谋的阴谋。

和发生在这个国家、发生在人们根本想不到的地方的事

情相比，古希腊人和罗马人的放荡性生活也是平淡的。(Adorno et al,1950：255-257)

不用说，在这个国家的不同地区，在社会的不同部门，对这些题器的阐释会有所不同。但主要的论点是：即使我们假定，在F量表建立之时，在F量表应答的基础上，我们有关权威主义的推论是有效的（我们不会探讨这个问题），我们也有把握地说，今天以F量表的原始形式来使用它，就等于测量一些事物，它们的性质和"权威主义"的最初概念毫无关系。因此，即使当代的研究者有意接纳阿多诺(Theodor W. Adorno)等人关于权威主义的表述和定义，我们也建议他（或她）使用修订的F量表或设计一个新的量表。

## 反射指标和构成指标

有学者区分了"反射指标"和"构成指标"（例如，Bagozzi & Fornell,1982）。前者也称为"反射因子"(Costner,1969)或"结果指标"(Blalock,1971)，它们是一类指标，我们把它们看做是所讨论的建构的效应。构成指标也称为"生产因子"(Costner,1969)或"原因指标"(Blalock,1971)，它们是另一类指标，我们把它们看做是所讨论的建构的原因。[①]

在有关建构验证的大多数研究中，指标都被看做是反射指标。有些学者，比较著名的有布莱洛克(Blalock, 1971)、科斯特纳(Costner, 1969, 1971)、豪泽(Hauser, 1972)、豪泽和戈德伯格(Hauser & Goldberger, 1971)、海斯(Heise, 1974)、雅各布森(Jacobson,1973)、乔雷斯考格和戈德伯格(Jöreskog & Goldberger, 1975)，也注意到在一些情境中，把指标看做是反射指标而不是构成指标，会更有意义。当然，这种选择并不是随意的；它依赖于有关建构的理论表述。以"社会经济地位"(SES)为例子，豪泽指出，虽然我们常常把受教育程度、收入等看做是SES的反射指标，但是，把它们看做是SES的构成指标，会更有意义(Hauser,1972)。当指标构成控制变量，并假定会影响不可观察的变量（例如，我们说，控制会影响焦虑、动机、挫折，参见后面的章节）时，很清楚，它们就是构成指标。

---

① 在这么早的阶段，我们不会探讨有关因果律的争论。在检验因果模型的语境下，我们将考察这些争论（参见第24章）。

## 模型图

模型图刻画了建构及其指标之间的关系,也刻画了各个建构之间的关系,它们的用途在于让我们一看就可以看出它们所推演的理论表述。卸掉词汇的包袱之后,一个模型图常常会揭示出理论表述中的缺陷、歧义和前后矛盾,它们原本是不易察觉的或者是不太明显的。因此,只要是探讨理论表述的时候,无论是自己的理论表述,还是他人提出的理论表述,我们都建议大家培养使用这类模型图的习惯。在这里,我们仅介绍一些惯例,在绘制模型图时,大多数学者会遵循它们。在后面的章节中,我们将介绍和讨论其他的方面。

不可观察的变量(潜变量)用圆来表示,可观察的变量(显变量)以长方形来表示。单向箭头表示因果关系的方向,从我们看做是因的变量(或自变量)指向我们看做是果的变量(或因变量)。变量间的相关关系则用双箭头曲线表示。

图 4.1 显示了上述观点以及与两类指标有关的观点。先来看一下图 4.1(a),可见,$X$ 和 $Y$ 是潜变量,$Z$ 和 $W$ 是显变量。$X$ 也设定为 $Z$ 的一个因,$Y$ 是 $W$ 的一个因。在建构验证的语境下,$X$ 和 $Y$ 是两个建构(例如,智力和动机、焦虑和攻击),前者由 $Z$ 来反映,后者则由 $W$ 来反映。最后,我们认为,$X$ 和 $Y$ 相关。

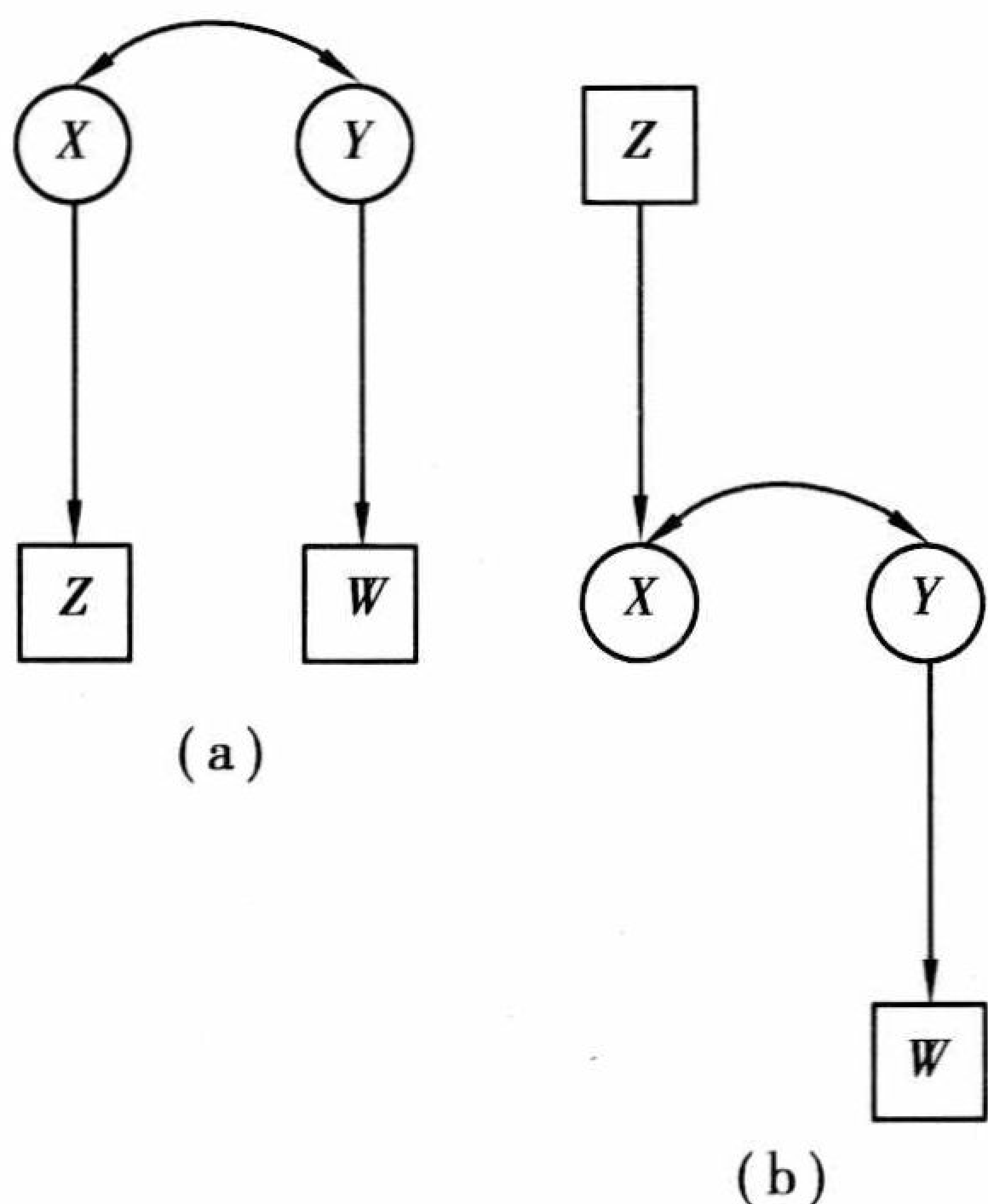



图 4.1

转到图 4.1(b),可见,像图 4.1(a)一样,$X$ 和 $Y$ 是不可观察的相关变量,$W$ 是 $Y$ 的一个反射指标。但与图 4.1(a)不同的是,在图 4.1(b)中,我们把 $Z$ 看做是 $X$ 的一个构成指标。

## 单指标和复指标

如图 4.1 所示,利用单指标来测量一个建构,总会遭遇各种难以克服的问题,因为我们不可能识别和分离该指标的变化度的不同来源。一般而言,一个指标的变化度由两个主要成分构成:系统性方差和非系统性方差。非系统性方差估计的有关问题也称为"随机误差方差",将在"信度"这个标题下加以讨论(参见第 5 章)。系统性方差可能有多种来源,包括该指标所反映的潜变量、其他潜变量、所使用的测量方法(例如,访谈、多选题器)和系统性误差(例如,选项集、社会赞许性)。

我们将采用一个简单的例子,来说明使用单指标时所面临的、几乎绝望的情境;我们也会揭示在这个问题的求解过程中所做的(隐含或外显的)不切实际的假定。假如我们想研究两个建构 $X$ 和 $Y$(比如焦虑和成就、自尊和工作满意度、挫折和攻击)之间的关系。依据 $X$ 和 $Y$ 是什么,以及我们的理论表述,一个假设可能是:这两者之间正相关或负相关。另一个可能的假设是:$X$ 影响 $Y$,或 $Y$ 影响 $X$,或两者相互影响。就当前的目的而言,让我们专注于"$X$ 和 $Y$ 相关"这个假设。在社会行为研究中,最普遍的做法是,每一个建构都使用一个单指标(或测量),然后计算它们之间的相关系数(参见第 17 章)。例如,就上述建构而言,我们可以计算自尊的自我评估和工作满意度的自我评估之间的相关系数。这种类型的设计如图 4.2(a)所示:$X'$和 $Y'$分别是 $X$ 和 $Y$ 的反射指标,$e_{X'}$ 和 $e_{Y'}$ 分别是 $X'$和 $Y'$的测量误差。

由于 $X$ 和 $Y$ 是不可观察的,因此,我们需要在它们推定指标之间的关系的基础上,来推论它们两者之间的关系。换句话说,$X'$和 $Y'$之间的相关系数,被视为代表了 $X$ 和 $Y$ 之间的相关系数。注意,这个推论的基础是一个隐含或外显的假定:每个指标与它所推定反映的建构之间相等。在图 4.2(a)中,这个假定表现在两个方面:一是从建构指向各自指标的箭头上的系数等于 1.00;二是由 $e$ 出发的箭头上的系数等于 0.00,表示这些测量不存在误差。

如果我们不作出这个假定(或其他有关建构及其指标间关系

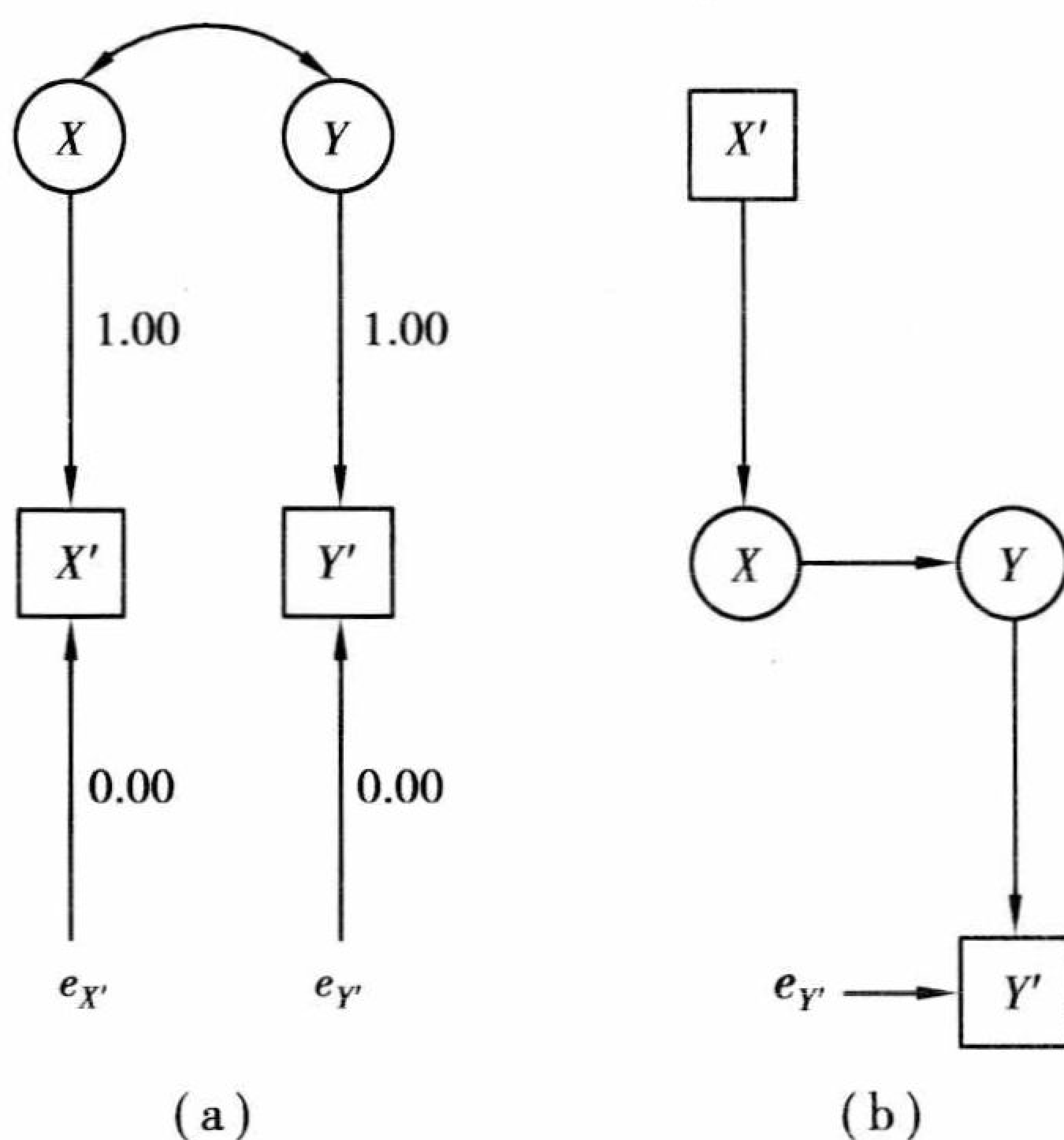



图 4.2

的假定),那么我们就要面对所谓的"欠判定模型"。它的基本含义是,我们没有充足的信息来求解模型中的未知系数(参见第 24 章)。就当前的目的而言,我们应注意到,如果我们删除图 4.2(a)中箭头上的系数,那么我们会面临使用单一信息(X'和 Y'之间的相关系数)来求解三个未知的系数(X 和 Y 之间的相关系系数、X 和 X'之间的系数、Y 和 Y'之间的系数),这是一项不可能完成的任务。

如第 5 章所示,有很多方法可以估计测量误差。使用这些估值,我们就可以在校正测量误差之后(第 5 章的"衰减校正"一节也将讨论这一点),求解 X 和 Y 之间的相关系数。但是请注意,当我们采用这种方法时,我们所参照的是图 4.2(a)所示的一种设计,除了建构所解释的变化度之外,我们还必须假定:余项完全是由随机测量误差造成的。在许多情形下,这个假定都会受到高度质疑。

在前面的例子中,我们使用的是反射指标。为有利于教学,我们在这里也讲解一个使用构成指标的设计例子。这个例子如图 4.2(b)所示,我们把 X'看做是 X 的一个构成指标,Y'是 Y 的一个反射指标。而且,为了更好地举例说明,我们假设不可观察变量 X 是不可观察变量 Y 的一个原因。

这个研究设计的示例是:假定 X'是一个影响心理建构(X)的

控制变量,[①]假设这个心理建构影响不可观察的变量 $Y$。这样,如果 $X$ 是焦虑,$Y$ 是学习,则 $X'$ 代表了一种引发焦虑或不同焦虑水平的控制。或者,如果 $X$ 是挫折,$Y$ 是攻击,则 $X'$ 是旨在引发挫折的一种控制。主要的论点是,研究兴趣不是控制($X'$)对 $Y$ 的影响(由 $Y'$ 来测量),而是建构($X$)对 $Y$ 的影响。[②]

和图 4.2(a)所示的一样,模型(b)也是欠判定的。这里也只有单一的信息,即 $X'$ 和 $Y'$ 之间的关系,在此基础上,我们无法估计出未知的系数。

我们希望上述讨论可以让大家相信,在社会行为研究的大多数领域里,当我们诉诸使用单指标,它就会把我们带入一种无法自圆其说的境地,以及随之而来的不切实际的假定。现在,我们转向探讨复指标的使用,这是旨在解决与使用单指标所伴生的一部分问题的行动的一个进程。

近年来,复指标设计的分析与概念有了重大进展。在起步阶段,我们的目标不是讨论这些发展,而只是在直觉的水平上,介绍和使用复指标有关的一些基本观念。在随后章节(例如,第 5、13、23 和 24 章)的多个语境下,对这里所介绍的观念会有详细讲解;处理复指标设计的、相关的方法论和实质性研究,有关它们的参考文献也会出现在这些章节中。

这里所使用的"复指标"一词,是指几个不同的可观察变量(或测量),我们相信它们都是同一个不可观察变量(或建构)的构成指标或反射指标。在当前的语境下,我们将基于多个题器的应答所形成的一个总分,看做是单指标。下面用一个例子来说明我们的意思。父亲的受教育程度、母亲的受教育程度、家庭收入、父亲的职业等,都可用作社会经济地位(SES)的复指标。但是,如果我们把这些指标合并成 SES 的一个指数,那么,在这种情况下,它毫无疑问就是一个单指标。

带着上述评注,让我们转到一个有关复指标设计的非常简单的例子,如图 4.3 所示。请注意,这个模型由两个建构组成:$X$ 和 $Y$,每个建构都由两个反射指标来测量。因此,$X$ 或许是创造力($X_1$

---

① $X'$ 并不一定是一个控制变量。参考前面的一个例子,$X'$ 可以是受教育程度,$X$ 则是 SES。

② 在第 8 章的"操作定义"一节,我们将讨论控制(或测量)与概念之间的关系。在第 12 章的"非意向效应"一节,我们将讨论图 4.2(b)所代表的设计的有关问题。

和 $X_2$ 是创造力的两个测量),$Y$ 则或许是智力($Y_1$ 和 $Y_2$ 是智力的两个测量)。

图 4.3 与图 4.2(a)相似,在这两张图中,研究兴趣都是两个建构 $X$ 和 $Y$ 之间的关系。但在图 4.2(a)中,每个建构使用的是单指标,在图 4.3 中,每个建构则使用了两个指标。如前所述,使用单指标来估计 $X$ 和 $Y$ 之间的关系,会碰到严重的困难,因为在大多数情形下,它的基础是站不住脚的假定。在图 4.3 中,$X$ 和 $Y$ 之间关系的估计,也是基于一组假定;我们在这里不会展开讨论,但要注意到,举一个例子来说,我们假定可观察变量的误差(例如,图 4.3 中的 $e$)之间不存在相关关系(它们之间没有连线)。

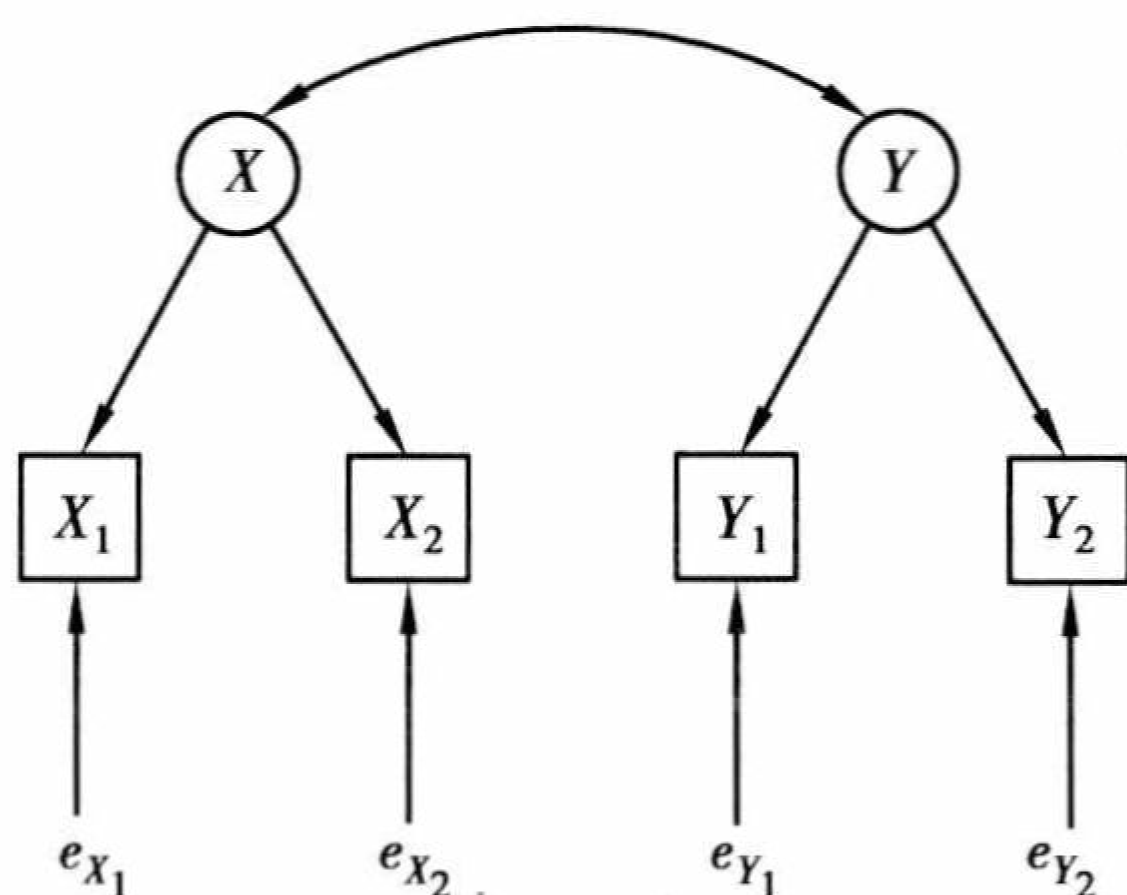



图 4.3

现在,我们要做的是比较图 4.2(a)与图 4.3,看看它们用于估计未知系数的信息的数量。在图 4.2(a)的设计中,可见,只有一个信息(例如,两个指标 $X'$ 和 $Y'$ 之间的相关系数),需要估计三个未知的系数;在图 4.3 的设计中,共有 6 个信息(当 4 个指标存在相关关系时,一共有 6 个相关系数,每两个指标计算一个相关系数),可以用来估计 5 个未知的系数。这样,我们可以说,在图 4.3 所示的模型是"超判定"的(参见第 24 章)。假如图 4.3 中的每个潜变量都使用了三个指标的话,和每个潜变量只有两个指标的情形相比,估计未知系数时,假定的限制更会少一些。

## 建构验证的方法

建构验证是一项永无止境的工作。在一个建构的各种指标的应答(或状态)的基础上,所作出的推论的可信度的得与失,取决于所累积的证据(包括所研究的建构)的性质和质量。因为检验包括

建构的假设,会对建构的验证产生影响,很明显,验证的方法只受研究者的想象力和敏锐性的限制,只受所研究的建构有关的理论表述和期望的限制。

方便起见,我们将在三个标题下讨论建构验证:(a)逻辑分析;(b)内结构分析;(c)跨结构分析。

## 逻辑分析

很明显,逻辑分析应该是研究活动的任何方面的一部分。在建构验证的语境下,逻辑分析的主要目的是形成各种竞争性假设,作为将要测量的建构、建构之间的关系等的备择解释。建构的定义、题器内容、测量方法以及使用测量的特定条件、给被访者的指导语和计分程序都是竞争性假设的来源。这些方面(以及尚未提及的其他方面)相互关联。而且,一个既定方面的相干性可能程度不同,这取决于使用这个测量的既定研究的特定条件。

"内容的逻辑分析无法反驳一个效度诉求"(Cronbach,1971:475),尽管这句话成立,但是,有时候,批判思维也足以让大家怀疑一个测量的效度,甚至让大家拒绝这个测量。正如克伦巴赫所说:"当有人指出,对听力障碍的儿童而言,听力问题会让我们对拼写测验的平常阐释变得不恰当时,不会有人还会如此迂腐、坚持索要'证明'"(Cronbach,1971:475)。

克伦巴赫的例子十分明显,引起了大家的共鸣。因此,我们想讲解一个研究例子,我们相信,这个例子也十分明显、达到不需要经验"证明"的地步,但很多研究者似乎会有别样的想法。"贝姆性别角色量表"(BSRI,Bem,1974)可能是当前测量男性气质和女性气质的最常用工具。就当前的目的而言,我们应注意到,BSRI 由一组形容词(例如,决断、上进、热情)构成,我们要求被访者在这些方面给自己评分。我们把"男性"特质和"女性"特质的评分分别加总,就得到男性气质分数和女性气质分数。贝姆认为,BSRI 中的男性特质和女性特质都是正面的或受人欢迎的。因此,为了检查被访者的应答是否存在社会赞许性,贝姆认为"在性别方面完全中立的[量表]",其中,一半题器"在价值上是正面的,另一半则是负面的"(Bem,1974:156)。

现在,贝姆声称是男性特质的一些特质,看起来像是正面的,但有些女性特质,不管怎么想象,都不能认为是正面的。"轻信"和

"阿谀奉承"就是这些特质中两个明显的例子。我们认为,在承认"这些特质不是正面的"之前,如果还坚持要求证明,那我们一定是迂腐的(在前面克伦巴赫所说的意义上)。不过,事情似乎是:只有证明这些特质是负面的、有些特质,甚至比贝姆的"中性"负面特质还负面之后(例如,Pedhazur & Tetenbaum,1979),一些研究者才会质疑它们的使用,然后,贝姆才把它们从 BSRI 短表中删除。(有关这个论点,参见 Beere,1983: 122-123)

附带说一下,对贝姆选择特质的方法进行一次逻辑分析,就足够揭示这种方法如何导致选出负面的特质(参见 Pedhazur & Tetenbaum,1979)。除了构建验证的逻辑分析之外,我们这里想强调的是,在我们综述的、使用 BSRI 进行研究的学者中,有相当数量的人似乎接受了贝姆的主张,即 BSRI 的男性气质和女性气质特质都是正面的、社会赞许的。一些女性特质可能是负面的、社会厌恶的,这种可能性,即使是一个暗示,我们也未曾碰到他们提及过。甚至相反,我们碰到了一些学者,他们毫不犹豫地宣称,BSRI 只包含社会赞许的特质。例如,凯利(Jeffrey A. Kelly)等人认为:"贝姆……在她的表述中只评估了男性气质和女性气质中正面的、社会赞许的成分。"(Kelly et al. ,1977: 1185)正因为如此,这些学者进行研究的唯一目的就是确定使用社会厌恶的特质(除了使用 BSRI 之外)的效应!尤其令人不安的是下列事实:甚至在同一个期刊中,这是大多数围绕 BSRI 的研究和争论发表的地方,主张"BSRI 仅由正面的特质构成"的论文仍旧得到发表。(最近的一些例子,参见 Larsen & Seidman,1986;Paulhus & Martin,1988)

现在,让我们就上述建构验证的逻辑分析的一些方面,进行简短的讨论,并举例加以说明。

### 建构的定义

逻辑分析最重要(肯定是首要的)的方面,或许是考察建构的定义。对定义在科学研究中的性质和作用,科学哲学家花费了大量的精力(参见第 8 章及其参考文献)。就当前的目的而言,我们想强调,考察建构的定义,对确定它是否有歧义、是否同义反复、是否逻辑一致、是否和建构所嵌入的理论结构一致等,十分重要。

众所周知,在一定的程度上,社会行为科学家会使用相同的建构来表示不同的事物。最好的一个例子是使用"态度"这个建构。

字面上,存在各种各样的态度定义,它们外显或隐含地源自各种理论取向。例如,有些态度的一般定义仅涉及评估成分,另一些定义也涉及认知成分和意志成分。一些研究者区分了一些建构,诸如“态度”“信念”“意见”和“价值”,另一些人则不加区分地使用它们。(有关著作,参见 Fishbein,1967;Greenwald et al. ,1968;Page,1980)

下面我们将举一个例子来说明,考察一般态度的定义,以及所研究的特定态度的定义,十分重要。假定一个人正在描述如何构建一个“自由主义—保守主义”的测量。自不必说,他首先必须界定这些建构。例如,他所关注的是政治自由主义,还是经济自由主义,还是两者皆有?另一个问题可能是,我们是在意识形态的层次上,还是在实用主义的层次上来评估自由主义—保守主义。例如,就政治信念而言,弗里(Lloyd A. Free)和坎特里尔(Hadley Cantril)发现,大多数美国人在意识形态上是保守主义的,在实用主义上则是自由主义的(Free & Cantril,1967)。

不过,我们或许还需要考察这个定义的另一个方面,即自由主义和保守主义之间的关系。有些研究者把自由主义—保守主义看做是一个两极谱系。它的基本含义是,高自由主义暗含着低保守主义,反之亦然。其他研究者则认为,自由主义和保守主义彼此独立。这样看来,一个人可能在自由主义和保守主义上都高,或者在两者上都低。(有关这个问题的详尽论述,参见 Kerlinger,1984)

更不用说,对建构定义的批判性评估有一个先决条件:了解和所研究的建构有关的理论和研究发现。对建构定义(以及建构验证的其他方面)采取批判性立场的一个好途径是,对特定领域内的建构测量的综述进行研究。致力于测量综述的出版物中,比较有名的有《心理测量年鉴》,目前由内布拉斯加州大学出版社出版。有些期刊(例如,《教育与心理测量》《咨询与临床心理学杂志》)也会定期刊载一些综述。

## 题器内容

研究者和测验设计者发现,我们可以方便地把一组题器称做一个“样本”,它来自作为一个既定建构的各种指标的题器库。在大多数情形下,这个题器库并不是在一个总体的意义上存在的,即我们不可能列举它的元素,并进行抽样。这样的话,我们如何设计

作为一个建构指标的题器呢?不论是在题器的设计阶段,还是从可用题器或其他指标中选择题器,建构的定义就是最重要的指南。同理,评估一个现存量表的题器恰当性时,我们首先要考察它们是否与建构的定义一致。①

## 测量程序

所谓"测量程序"是指测量的一般方法(例如,访谈、总和评分量表、语义差异、投射技术)、这些方法的特征(例如,给被访者的指导语、题器顺序、题器措辞),以及实施条件(例如,被访者的匿名性保证,该测量与推定是其他建构的测量一起实施)。

所得分数受到所使用的特定测量程序影响越大,对测量效度的负面效应就越大。测量程序的一些方面,或多或少会影响被访者的应答,影响的幅度取决于使用该测量的研究本身的特征。例如,在关于"敏感"问题的态度研究中,是否保证被访者的匿名性,就变得非常重要。同样,当研究关注废除种族歧视的政策时,访谈者的种族就变得非常关键。

这里,我们想强调的论点是:我们必须在研究的总体目标和设计的背景下,依据所研究的测量的特定属性,来审视测量程序。我们将在第6章讲解社会行为研究中所采用的一些主要测量方法。在本章的后面章节,我们将介绍如何使用多个方法来测量相同的建构,以便把源自特定方法的方差和源自建构的方差分离开来。眼下,我们只举几个例子,考察测量程序以及随之而来的竞争性假设。

在本章的前面,我们曾简要介绍和讨论过F量表(或后来众所周知的"权威量表")。F量表是总和评分量表(常常也称为"李克特量表")的一个例子(参见第6章)。就当前的目的而言,我们应指出,在总和评分量表上,一般要求被访者指明自己同意或不同意一组陈述的程度;然后,我们把答案加总,得到一个总分。为了消除或降低选项集效应,一般建议,量表应同时包含正面和负面措辞的陈述。

F量表发表后不久,有些研究者指出,它的所有陈述都以相同

---

① 请注意,这里的讲解仅限于建构验证的逻辑分析方面。评估题器恰当性的其他方法,将在后面的章节(特别是"内结构分析"一节)加以讲解。

的方向措辞;他们争辩说,高分并不能反映高权威主义,而是反映部分被访者的赞同或"答是"倾向。("答是"或"答否"是一个人格变量,有关讨论,参见 Couch & Keniston,1960)这样,对测量方法的一个特点的逻辑分析,引起了对 F 量表的建构效度的质疑。(参阅 Bass,1955;Christie et al. ,1958;Gage et al. ,1957)

我们来看一下"威尔逊-帕特森态度量表"(WPAI,Wilson,1975),作为对测量程序进行逻辑分析的另一个例子。它假定是保守主义的一个测量,尽管据称它也能在自由主义和其他一些建构(参见"计分"一节)上产生一个分数。下面是给被访者的指导语:

> 你喜爱或相信下列哪一个?
>
> 圈"是"或"否"。如果您实在不清楚,圈"?"。答案没有对与错之分;请勿讨论,给出您的第一反应即可。请回答每一道题。

这些指导语之后,跟着 50 个参照物(例如,绞死小偷、太空旅行、比基尼)。

请注意,被访者无论是喜爱,还是相信一个既定参照物,都会给出"是"的答案;反之,则给出"否"的答案。对一些参照物而言,虽然信念和评估相互关联,但将两者等同,则可能是歧义的一个来源。更多的歧义来源是题器格式和应答模式。在对 WPAI 的综述中,佩达泽对这些问题进行了评论。他认为:

> 例如,考量一下"抽大麻"这个题器,"是"的答案可能是指:支持抽大麻的合法化,但不支持抽大麻;同时支持合法化和抽大麻;支持抽大麻,尽管(或由于)法律禁止。其他的含义,也还是可以想象的。同理,对同一个题器的"否"的答案也可能具有多个含义。(Pedhazur,1978: 1151)

## 计分程序

计分程序也会影响基于它们之上所作的推论的效度。① 因此,对计分程序的评判性评估,是逻辑分析的一个重要方面。特别是当我们采用一个测量来探测一个多维建构,或是所引起的应答需要我们进行分类和编码(例如,投射技术)时,有关计分的问题可能

① 计分程序也会影响一个测量的信度。信度及其与效度的关系将在第 5 章中加以讨论。

会变得十分复杂。

甚至当计分看起来似乎是直截了当的时候,复杂的情形也会出现。例如,就计分而言,成就的多项选择测验似乎并不存在什么问题。多数人将每个正确答案赋值为1,每个错误答案赋值为0,然后,将正确答案的个数加起来,得到总分。然而,可能还需要作出各种决策。例如,可能有必要决定是否对猜测进行校正,以何种方式校正(参见 Nunnally,1978: 642-65);或者,是否对不同的题器赋值不同的权重(例如,Nunnally,1978: 296-297;Stanley & Wang,1970)。

我们的目的不是在技术层面上论述计分程序,而是要说明,在这方面常识也大有裨益。下面,我们首先介绍一个题器的应答计分,然后再探讨在两个或多个题器的基础上形成一个复合分的问题。

一个题器不过是一个建构的一个指标,如果我们不考量这个建构,那么我们就无法作出有关这个题器计分的决策。我们再拿 WAPI 来做一个例子(参见前面),这次我们考察的是题器计分的一些问题。如前所述,WAPI 被推定是一个由 50 个参照物(例如,校服、主日学校、向国旗敬礼、比基尼、留胡须的男人、漫画)组成的测量,它测量保守主义、要求被访者用圈“是”“否”或“?”的方式作答。对一个“保守主义”参照物,一个“是”的答案赋值为 2,一个“否”的答案赋值为0;相反,对一个“自由主义”参照物,一个“是”的答案赋值为 0,“否”的答案则赋值为 2。对这两种参照物而言,一个“?”的答案赋值为 1。[①] 我们把 50 个题器的得分加起来,就得到一个“保守主义—自由主义”谱系上的分数,高分表示倾向于保守主义,低分表示倾向于自由主义。

请注意,威尔逊把自由主义—保守主义看做是一个两极谱系,这是他的计分程序的概念基础。[②] 在这里,我们暂不关注这个概念的优点或缺点,而是关注它所带来的令人质疑(甚至是奇怪)的计分程序,依据的是测量所使用的参照物和应答方式。例如,看一看

---

① 下面是威尔逊关于该类别的意义和计分的表述:“我们用‘?’这个类别表示‘不理解’‘中立’或‘不关心’;对这些应答中的任何一个,赋值一个中间分,看起来是合理的”。(Wilson,1975: 18)我们想让大家判断威尔逊这个假定的“合理性”。而且,请思考一下,他在给出参照物的类型和对应答方式提出要求时,使用的是“实在不清楚”(参见前面的指导语)一词。

② 有关这个论点的讨论和参考文献,参见“建构的定义”一节。

下列参照物:春宫图、吸大麻、嬉皮士、脱衣舞和裸体游泳。对这些参照物中的每一个,一个"否"的答案赋值为2,一个"是"的答案赋值为0。为了方便讨论,我们假定对这些参照物的"否"的答案反映了保守主义的态度。但是,为了让答案能够在通向"自由主义"一极的方向上得到计分,为什么一定要在这些参照物中的每一个上面回答"是"呢? 正如佩达泽所言:"期望自由主义者对这些参照物展现宽容,是一码事;期望他们说喜爱或相信它们,则完全是另一码事。"(Pedhazur,1978: 1151)

现在,我们转而探讨有关复合分的问题。如前所述,有关建构的大多数测量都是由多个题器组成的,因为单个题器涵盖一个建构之领域的情形,即使可能的话,也极其罕见。把对一组既定题器的应答合并成一个复合分,这种做法是否有意义,我们将在下一节("内结构分析")探讨一些方法来加以确定。但目前我们只想强调,合并一组题器上的得分,判断这种做法是否有意义,大可不必诉诸各种复杂的分析。经常发生的情形是,在求得复合分的过程中,人们完全无视基本的测量原则,甚至是常识。正如邓肯(Duncan,1984:227)所言:

> 在化学实验室中,我们知道,应谨慎对待物质的"组合"。但据我所知,某种相似的、信息的"组合"也具有一个类似于实验员所面对的属性,这还没有得到广泛的认可。

下面举个例子来说明这种现象。先前,我们提到了阿莫尔曾对来自《科尔曼报告》的数据进行了二次分析(Armor,1972)。作为二次分析的一个方面,阿莫尔的兴趣是研究教师对一些选定问题的应答和学生成绩之间的关系。对前者而言,阿莫尔选择了6个问题来问教师,他用这些答案,计算一个复合指数。阿莫尔(Armor,1972: 228)的表述是:

> 如果一个教师:①在重新选择前面仍会坚持选择教书;②不想换学校;③对(学校的)民族构成表示没有偏好;④对(学校的)种族构成表示没有偏好;⑤希望一直教书、直到退休;⑥喜欢看到一个黑人学生上白人为主的大学,那么,上述每一种情形,他都将得1分,由此我们推导出一个指数。

这段引文是对题器及其选项的复述。例如,阿莫尔所用的第一个题器如下:

假定您能回到过去,重新上大学。以您目前的阅历来说,您还会进入教书这一行吗?

(A)肯定会

(B)可能会

(C)不确定

(D)可能不会

(E)肯定不会

(Coleman et al. ,1966: 679;完整的"教师问卷"以及研究中所采用的其他测量,都刊载在《科尔曼报告》中。)

这里,我们所关注的是复合分,不是阿莫尔的题器计分策略,尽管我们对后者持保留意见。例如,和上面这个题器的计分有联系,阿莫尔把"肯定会"当成一个类别,把其余的选项答案勉强凑成另一个类别,我们不明白这样做的道理何在。有关阿莫尔的题器计分的这方面以及其他一些方面的问题,我们请大家进行思考(Armor,1972)。

下面是阿莫尔有关他的"指数"概念的全部内容:"我们预设的概念是,对教书和学校的强承诺、对民族构成的一种宽容态度(不反对任何的种族或民族合校)应是一个好老师的态度组成部分。"(Armor,1972: 179-180)即使忽视这段话的歧义性,也不讨论"教书态度甚至没有暗含到主要的方面"这个问题,我们也不得不质疑:将上述三个方面看做是一个建构的各个部分,是否有道理、有意义。例如,对少数民族群体的宽容,可能与教书态度有关。当然,这并不意味着,对少数民族群体的宽容是"教书态度"这个建构的一部分。

即使我们不会质疑阿莫尔对教书和学校承诺的测量(这是后面的质疑)诉求,我们也不得不质疑:它们为什么有必然联系,更不用说,它们是同一个建构的两个部分。对一个教师而言,他对教书具有高度承诺,但由于各种原因(例如,交通便利、声望),他希望调动到另一所学校,这难道是不可能的吗?实际上,相对比较容易碰到的典型情境是,一个教师因为对教书的承诺而寻求一次调动。

利用一个或两个指标来探测一个复杂的建构,在不深究这种令人怀疑的实践前提下,让我们对它们进行简短的考察。以假定反映教书承诺的两个指标(例如,在重新选择前面仍会坚持选择教书;希望一直教书,直到退休)为例,对这两个问题的肯定回答可能反映了教书承诺的缺失,支持这种观点的论证,还是相对比较容易

得出来的。在最低限度上,我们也必须接纳这样的观点:它们可能反映了其他事物(例如,评估录取、顺利完成学业、在其他更喜爱的行业中接受培训等机会;与自己孩子的假期同步的几个长假期的吸引力;工作保障;拒绝墨守成规)。阿莫尔对其他题器应答的阐释,我们也可以提出类似的质疑。

当阿莫尔使用这个指数来考察他所感兴趣的关系时,究竟会发生什么?在上述讨论的背景下,我们有兴趣来考察这个问题。阿莫尔(Armor,1972:180)的报告称:

> 我们发现……好教师的态度指数与学生平均成绩之间呈反向的关系。事实上,就全美总体而言,教师态度和学生成绩之间的积矩相关系数是 -0.42。学生成绩越高,教师的宽容和承诺越低(以我们的指数所测量)。一种可能是,我们关于好教书态度的观点是错误的,但我们更愿意相信,由于某些原因(!),该指数不适于评估这些态度。

在对《科尔曼报告》数据的后续分析中,阿莫尔明智地决定不再使用该指数。我们的讲解旨在表明,他为什么从一开始就注定无法形成这样一个指数。

在一个建构的验证过程中,批判性思维、理论、测量、设计和分析的知识之间相互影响。我们希望,对逻辑分析的这些讨论能够说明这种相互影响的重要性。现在,我们转向建构验证过程的另一个方面,即"内结构分析"。

## 内结构分析

在逻辑分析的层次上,我们已经在上一节探讨了涉及指标的选择、计分方式、分数的组合方式等问题。在这一节,我们将致力于描述各种分析方法,它们的目标是评估反映同一个建构的一组指标的效度。从建构验证的视野来看,这些讲解的主要篇幅构成了对因子分析方法的直观描述。第 22 和 23 章的主要内容是因子分析。根据大家的背景、特定目的和需要,把这两章和下面的内容合起来一起学习,可能是有益的,甚至是必要的。

请看图 4.4,其中,$C$ 表示一个建构(例如,智力、攻击、态度),$X$ 表示指标(例如,智力的测量、打人或摔物、骂脏话、测量态度的题器),$e$ 表示误差或一个既定指标的特有方面。如图所示,每个指标

都有两个构成成分:(a)一个源自该建构;(b)一个源自误差或该指标所独有的其他因子,其证据是 $e$ 之间缺乏连线。[①]

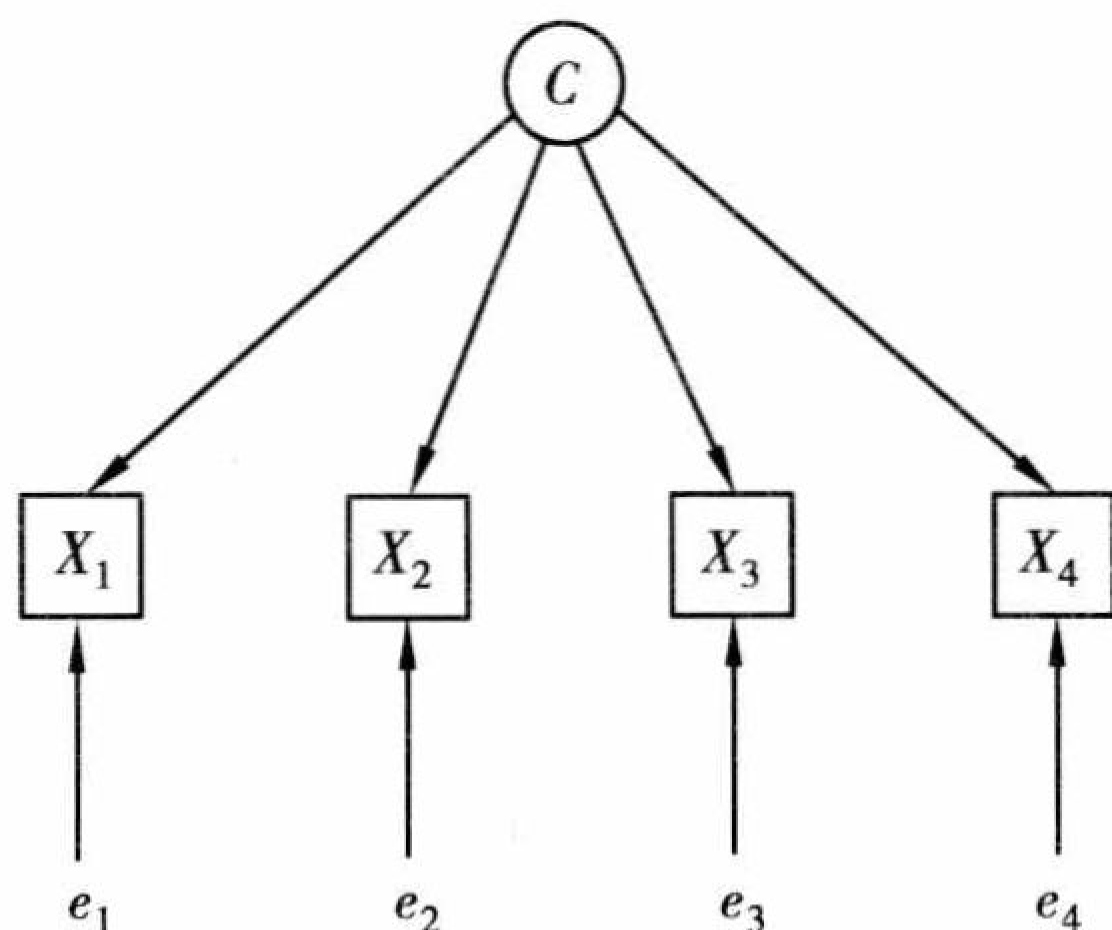



图 4.4

由此可见,在图 4.4 中,各指标之间的相关关系,可以归因于它们所反映的这个建构。因此,为了接纳这个模型的效度,因而也接纳这个建构的指标的效度,我们有必要(尽管并不充分)证明:该模型和数据一致。本质上,这意味着各指标之间关系能够被该模型合理地解释掉。[②]

在最低限度上,我们必须证明,一个建构的反射指标(例如,题器、子测验)"粘在一起",它们是同质的。自不必说,各种异质的指标不可能测量同一个事物,因此,将它们组合成一个复合指数,这种做法是荒诞的。

## 因子分析和内部结构

研究一组指标的内部结构,存在很多方法。最有效的方法或许是因子分析的一些衍生方法(这是第 22 和 23 章所关注的主题)。就当前的目的而言,我们只对这个方法作一个一般讲解,并说明它在建构验证过程中的潜在用途。

"因子分析"是指一系列旨在识别因子(或维度)的分析技术,这些因子隐藏在一组观察变量之间的关系背后。在当前的语境下,这些可观察的变量是各种指标(度量、题器),我们假定它们反映这个建构(例如,因子)。

---

① 我们有可能设定各种误差项之间相关的模型(参见第 23 章)。

② 各种备择模型也可能与该数据一致。这个事实,我们将在其他章节(特别是第 22、23 和 24 章)中加以讨论。

在大多数情形下,我们把因子分析应用于指标之间的相关系数。我们可以求得每个指标与一个因子之间关系的估值(称做"因子负荷")。宽泛地说,一个因子负荷是一个指标在因子上的权重,就像回归分析(参见第 17 和 18 章)中的标准化回归系数(β)一样。对某些因子分析的解来说,因子负荷是指标和因子之间的相关系数。在这种情况下,一个因子负荷可以在0(指标和因子间不相关)和正负1(指标和因子间完全相关)之间变化。

一般来说,因子负荷越大,因子对指标的影响越大。和相关系数的阐释一致(参见第 17 章),因子负荷的平方也是指一个既定指标的方差被这个因子所解释掉的比例。例如,0.4 的因子负荷意味着这个因子能够解释掉 0.16($0.4^2$)或 16%(0.16 * 100)的指标方差。

我们把探索性因子分析和证实性因子分析作了一个区分。宽泛地说,探索性因子分析所处理的问题是,我们需要多少个因子,才能解释一组指标间的关系,才能估计出因子负荷。顾名思义,证实性因子分析所关心的是参数估计和假设检验,如关于一组指标间关系背后的因子数的假设。

下面,我们将对这两种方法进行一般的讲解,并距离说明它们在建构验证中的用途,尝试厘清这两者的含义。在这之前,我们要做两点说明:第一,正如我们将在第 22 和 23 章所说的那样,关于这两种方法的内容还有很多,甚至多于我们在这里所暗示的;第二,我们十分清楚,大家很可能会觉得我们的概述令人费解,甚至令人沮丧。更糟糕的是,我们也知道,当我们对极端复杂的分析方法进行松松散散的描述时,我们冒着极大的风险:这可能会误导大家,让大家形成错误的概念。因此,在对因子分析形成一种意见之前,更不用说,在尝试阐释因子分析的结果之前,大家至少应该先学习第 22 和 23 章。

### 探索性因子分析

举例来说,假定一个研究者有兴趣构建一个"自我概念"的度量,他(或她)设计(或从文献中选择)了 20 个题器,并要求被访者在这些题器上对自己进行评级。再假定这个研究者想把这 20 个题器的评分相加,得到一个"自我概念"的总分。顺理成章的是,如果这个总分具有意义的话,基本的要求是这些题器"粘在一起",指向相同的维度。

当然,他可以研究一个题器与其他每一个题器之间的相关系数。不过,即使题器的数量相对较少时,他一般也难形成一个总印象。他需要考察和总结的相关系数的数量等于:

$$\frac{k(k-1)}{2}$$

其中,$k$ 表示题器的数量。这样的话,就这里所考察的 20 个题器而言,他就必须考察:[(20)(19)]/2 = 190 个相关系数。但如果他对相关系数矩阵进行因子分析,并且只保留一个因子(参见下面的讨论)的话,他就只需要考察 20 个因子负荷。

考察因子负荷的目的是确定,它们中的哪一些与这个因子存在有意义的相关关系。正如我们在后面几个章节(例如,第 9 和 15 章)中所讨论的那样,统计量(在当前的情形下,是相关系数)的"意义"取决于各种考量,其中,首要的考虑是研究的实质方面。在因子分析的应用中,研究者通常将大于 0.4 或 0.5 的负荷看做是有意义的。采用类似这样的准则,这个研究者就能够决定,只把那些负荷有意义的题器包括在量表中。

当他认为负荷满足"有意义"这个准则的题器数量不足时,他会加入新的题器,或修改那些负荷未达标的题器。一套新题器由负荷达标的题器和新加或修改的题器所组成,他对它再进行因子分析。这个过程或许会一直重复下去,直到这个研究者对这套题器感到满意为止,它构成"自我概念"的一个度量。

为了避免重复尝试,经常采用的做法是,从一个较大的题器库起步,量表中所用的题器可以从这个库中加以选择。对这个较大题器库的因子分析,可以帮助我们决定:保留哪些题器、删除哪些题器、修改哪些题器、拥有什么。参阅戈萨奇(Gorsuch,1983:第 17 章)、马拉迪(Marradi,1981)、泽勒和卡迈恩斯(Zeller & Carmines,1980:第 4 章)有关量表建构过程中运用因子分析的讨论,这对大家可能会有所帮助。

上述讨论必然建立在若干暗含的假定之上,现在,对其中的一个最重要假定,让我们作一点说明,并作简要讨论。正如前面多次提到的那样,我们必须首先定义一个建构,然后才可能去寻找相干的指标、设计或选择题器,等等。在前面的讨论中,我们曾假定"自我概念"是一个单维建构。现在,我们反过来假设它是一个多维度、多剖面的建构(例如,夏文森和布勒斯提出"学术自我概念""社

会自我概念”等,参阅 Shavelson & Bolus,1982),在这些情形下,我们会设计或选择一些题器,假定它们可以代表每一个剖面。当我们将所有题器的相关系数矩阵用于一个因子分析时,我们的期待是,旨在表示“自我概念”的不同剖面的题器,应在不同的因子上具有“有意义”的负荷。但是,如果我们把不同的剖面看做是一个普遍的“自我概念”的各个维度时,那么我们的期待是,反映这些维度的因子之间会存在相关关系。

前面的讨论中所暗含的另一个假定是,在因子分析中只会“浮现”一个有意义的因子。[①] 现在,假定一个研究者将“自我概念”定义为单维建构,有一个相关系数矩阵,它属于旨在测量这个建构的题器,当我们对它进行因子分析时,出现一个强烈的征兆,说明在这些题器的相关系数背后,存在两个相对独立的因子。面对这样的结果,这个研究者可能会坚持原来的概念,即“自我概念”是单维的,并得出结论说,开发量表的尝试失败了或只取得了部分成功。在这样的结论下,这个研究者可能会决定:只保留一些题器,它们只在其中的一个因子上具有“有意义”的负荷;他(或她)认为,这个因子与“自我概念”的定义一致。

如果有必要的话,他可以修改某些题器、设计或选择新的题器,这样,上述过程就会重复进行下去。假定他的确进行了重复,但多个有意义的因子再次“浮现”出来。原则上讲,没有什么可以阻止一个研究者不断努力去创建一个度量,他(或她)认为测量的是一个单维建构。顺理成章的是,如果不能创建一个和建构一致的度量,在不断的失败面前,这个研究者可能不得不诉诸其他行动方案。其他因素之外,这取决于这个研究者的信念、理论表述和意志,他(或她)可能会放弃创建这种度量的努力,转向不同的测量方法,找出不纯的变量,修改建构的定义,这里仅举出几种选项而已。

不论采取何种行动方案,研究者和公众都应当小心“物化陷阱”,它的典型表现是追问:一个特定是否真的测量了所研究的建构?(例如,韦克斯勒智力量表真的测量了智力吗?)这是一个错误的问题,因为它忽视了一个事实,即我们所处置的是抽象的,而不是一个对象,我们无法使用这个度量,看看它是否为这个对象“量

---

① 应该保留多少个因子的准则是什么?这个问题非常复杂,争议不断。有关讨论,参见第22章。

身定做”。

关于一个度量的一个有意义的、基本的问题是:它是否与它所想要测量的建构的定义一致?例如,在考量智力的一个度量时,我们有必要考察它是否与智力的定义一致,这是它想要勘察和反映的建构。就这个目的而言,因子分析可能是我们可以使用的、最强大的分析方法之一。事实上,智力测量的历史、围绕这些尝试的争论,是因子分析发展历史、它的推广应用历史、围绕它的争论历史的重要组成部分。

斯皮尔曼(Spearman,1904)奠定了因子分析的基础,他坚定不移地使用这种方法,为自己的两因子智力论提供支持。根据这个理论,一个公共(或一般)因子(g)是智力的各种度量的基础;一个独特(或特殊)因子(s)是每一个度量的基础。同样,瑟斯通(Thurstone,1947)认为,智力是由相对分化的心理能力所构成,他采用因子分析来开发一种基本智力的度量。因子分析在智力的理论发展及其测量的建构验证中的作用,一个典型的示例是吉尔福德(Joy P. Guilford)有关“智力结构”的广博研究。(心理测试发展的一个总结以及历史背景,参见 Guilford,1967)

在阅读上述讨论的过程中,大家可能会问:在建构定义和测量方法之间的差异面前,我们怎样才能决定哪一个是“正确”的呢?请注意,如前所述,一个建构本身是没有意义的。一个建构的意义和相干性源自它所嵌入的理论背景、语义网络——我们将在“跨结构分析”一节中返回这个主题。

总之,在建构一个度量的阶段,探索性因子分析可能很有价值。例如,一个研究者相信他(或她)已经建构了一个单维度量,但经过因子分析,他可能会发现,这是多维度量,在该量表上,一个单一的总分难以成立。反之,假定已经建构的是一个多维度量,一个研究者可能发现,它是单维的。

对一个想应用一个现有度量的人来说,因子分析也可能很有价值。对一个现有量表进行因子分析,可能是非常基本的或极其有用的,下面就是证明这一点的几个例子:

**1. 当关于度量的内部结构的信息不存在时(在社会行为研究中,这是十分平常的现象)。**在因子负荷的基础上,如果我们可能得出结论:一些题器不相干(即它们的负荷非常低),或者它们与量表所假定测量的建构定义甚至相矛盾(即它们

的负荷符号是"错误"的,它们在"错误"的因子或多个因子上具有"有意义"的负荷)。

2. **因子结构不同于这个度量的设计者所报告的结构,当情形看似这样,或被期待是这样时**。其他原因除外,出现这种情形的原因可能是,我们采访的被访者类型不同(例如,男性与女性,青年与老年)、源于历史的变迁(例如,战前与战后的条件)。

3. **当潜在的使用者对量表的因子分析的恰当性持有保留意见时**(例如,抽样和(或)样本规模、保留的因子数、因子的析取或转置方法)。我们将在第 22 章讨论和说明诸如此类的一些问题。

### 证实性因子分析

如前所述,证实性因子分析涉及参数估计和假设检验。在这一节中,我们的目的是有限的。我们想做的一切,就是要用非常宽泛的语言,描述证实性因子分析的用途是什么,它会带来什么。为了达到这样的目的,我们将会采用相同的实质示例,即"自我概念"的测量;在探索性因子分析中,我们已经列举过它。探索性因子分析和证实性因子分析之间的对照,我们留在第 22、23 章中进行详述,那时,我们将采用这里所举的实质性示例,把两种分析方法运用到同一套示例数据上。

这样的话,假定我们想对"自我概念"的各种指标(即题器)进行证实性因子分析。和探索性因子分析不同,很明显,在证实性因子分析中,有待检验的假设必须先行一步。举例来说,假定我们的假设是,"自我概念"由两个相关的维度(学术和社会)构成,每一个维度由三个指标所勘测,如图 4.5 所示。

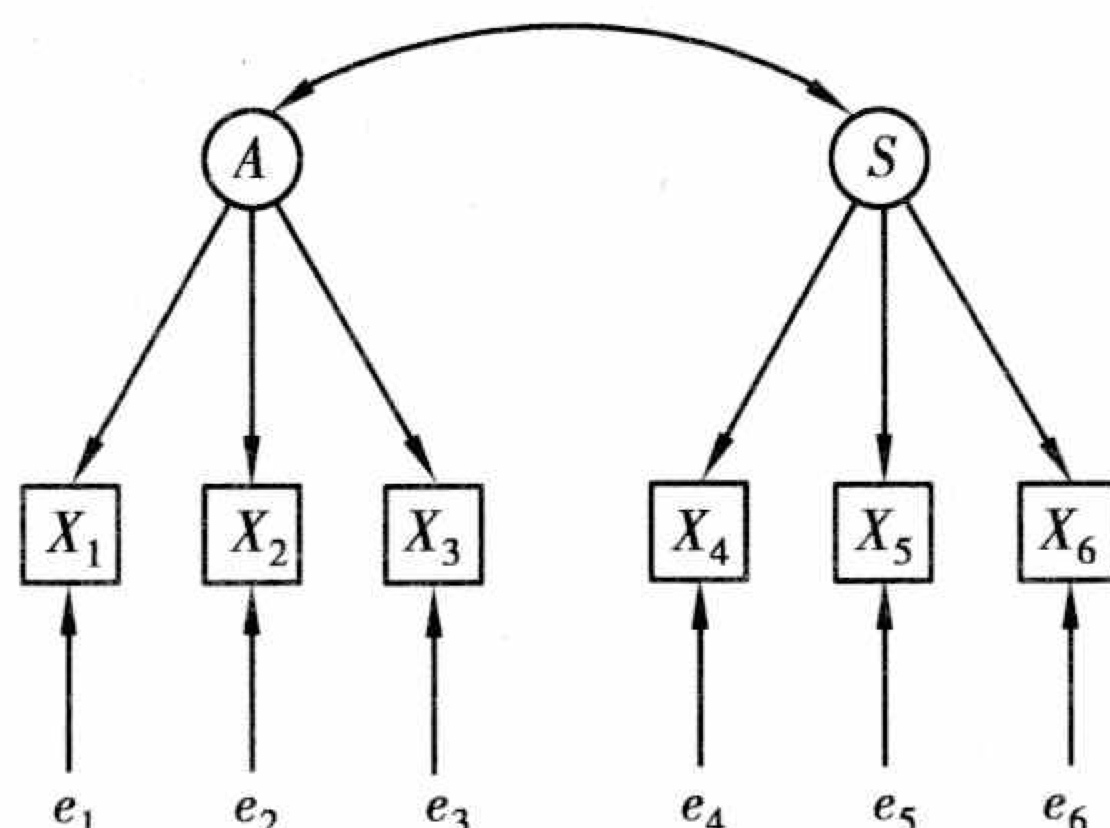



图 4.5

按照本章前面所描述的惯例,建构(因子、潜变量)用圆表示,指标(显变量)用方框表示。在图 4.5 中,$A$ =“学术自我概念”,$S$ =“社会自我概念”,$X$ 表示指标。弯曲的双向箭头表示我们认为“自我概念”的两个维度之间存在相关关系。请注意,依据这个模型,每个指标仅反映(具有负荷)一个因子,误差(即 $e$)之间也不存在相关关系。使用指标之间相关系数(或协方差),我们就可以采用证实性因子分析,估计出指标在因子上的权重、因子之间以及误差之间的相关系数(或协方差)。[①]

宽泛地讲,模型检验就是确定:利用参数估值,我们是否能够复制出或近似逼近所观察到的关系(相关系数或协方差)? 请看图 4.5,如果模型拟合数据的话,如利用(相乘)$X_1$ 和 $X_2$ 在 $A$ 上的权重,我们就有可能复制或近似逼近它们之间的相关系数,由这个模型可见,因为这是它们之间唯一的共同点。相比之下,如 $X_1$ 和 $X_4$ 之间唯一的共同点是,它们各自所反映的因子($A$ 和 $S$)是相关的。相应地,当我们试图复制这两个指标之间的相关系数时,我们就必须利用(相乘)因子之间的相关系数和指标在因子上的权重。

我们非常担心会带给大家一种错误的印象,即证实性因子分析的过程直截了当、不容争辩。但我们相信,在这里讨论让这个过程变得复杂、模糊的各种问题,也是不智之举,因为这些问题本身就很复杂,因此,我们把它们放到后面的章节中加以详细阐述。

在证实性因子分析过程中起主要作用,并在后面的章节中将要加以详述的问题包括:

(1)模型设定。当我们检验一个模型(例如,这里所考察的模型)时,我们必须假定:它的设定正确。

(2)假设检验的逻辑。严格来讲,我们不可能证实一个假设。我们所能做的一切,就是拒绝或无法拒绝假设。我们无法拒绝一个模型,这让我们得出一个结论:我们证实了这个模型。当我们认识到,可以证明有很多其他模型也可能拟合这个数据时,十分明显:这是用词不当。

(3)统计显著性与实质重要性之间的差别,以及样本规模、效应规模、显著性检验的统计力等相关问题。

虽然上述内容可能会使大家感到困惑或害怕,但我们保证,这

① 具体的算法和步骤,将在第 23 章中加以解释和说明。

并不是我们的本意。我们所面对的是一个两难困境：要么掩藏这种复杂性，冒着把大家引向错误使用和错误阐释的道路之上的风险；要么暗示这种复杂性，冒着让大家感到困惑和害怕的风险。我们强烈地感到，冒后一种风险更好。

无论大家对上面罗列的各种问题作何反应，我们的建议是，在这个阶段，大家不必担心它们。正如我们所言，在后面的章节中，我们会详加讨论。我们所想做的一切，就是让大家注意：为了能够成为使用证实性因子分析的明白人，大家就必须熟悉它们。

## 跨结构分析

在前一节，我们讨论并举例说明，为了确定一组指标的结构和它们所反映的建构之间是否一致，内结构分析是必要的。但是，就支持一个度量或一组指标的建构效度而言，来自内结构分析的证据是必要而非充分的证据。认识到这一点十分重要。原因在于，一个既定的内部结构可能与不同的建构定义一致。例如，当一个建构被界定为单维时，那么，就有必要证明：旨在反映这个建构的一个度量或一组指标的结构也是单维的。然而，这个证据并不能排除这种可能性：这是一个不同于研究者脑中所想勘测的单维建构。而且，这种结构也可能是源自我们所使用的测量方法所特有的方面（参见后面“趋同效度”和“判别效度”标题下有关“方法因子”的讨论）。

如前所述，一个建构的含义在于它与语义网络中其他建构的关系上。因此，归根结底，建构验证的基础是研究当前建构与理论框架中的其他建构（或变量）之间的关系。我们用“语义效度”这个标签来表示这种方法（Campbell，1960；Cronbach & Meehl，1955）。

跨结构分析相当于假设检验，其中，当前的建构是诸多变量之一。例如，假定我们关注两个建构（$X$ 和 $Y$），并采用复指标来测量每一个建构。按照上一节所讨论的程序，我们可以研究每个建构的一组指标的内部结构。不过，除此之外，我们还有必要进行跨结构分析。请注意，我们研究的是两个或多个建构的指标之间的关系，正是在这个普遍意义上，我们才使用“跨结构分析”这个术语。例如，我们可以采用一个理论框架，假设 $X$ 和 $Y$ 正相关，或者，$X$ 影响 $Y$，或者，$Y$ 影响 $X$。无论假设是哪一个，$X$ 和 $Y$ 都是不可观察的（潜）变量。因此，假定一些指标是当前建构的外显特征，我们通过

研究这些指标之间的关系,就可以进行假设检验。

假设得到支持,表示 $X$,$Y$ 或两者的指标(度量)效度得到支持。如果假设得不到支持(参见第 9 章对"假设检验"的讨论),并不一定表示 $X$ 的指标没有效度。假设得不到支持的其他阐释包括:(a)理论框架值得怀疑;(b)$Y$ 的指标效度值得怀疑;(c)研究设计或分析中存在缺陷。(有关负面证据的含义,参见 Cronbach & Meehl,1955: 295-296)

对建构验证而言,跨结构分析是非常重要的;我们必须时刻牢记,即使待检验的模型能很好拟合数据,还可能存在其他的解释和模型。例如,假定我们发现,一个模型由两个建构的复指标所构成,它拟合数据;我们认为是两个建构的事物,有可能是一个建构的两个不同剖面(正如和图 4.5 关联的那个例子所表明的)。

一组备择模型,它们同样很好地拟合了一个数据,我们如何从中进行选择? 对这个问题并没有简单的答案。(对这类问题的探讨、所推荐的解答的讨论,参见 Tesser & Krauss,1976)检验一个模型拟合与否,只是理论和可观察现象之间相互作用的一个方面。总的来说,一个理论,以及由此推导的模型越精细,当我们发现这个模型拟合数据时,我们对结论的信心就越大。不过,一个既定模型拟合一组既定的数据,这个发现并不构成一个确凿的证据:它就是"真正"的模型。从不同的理论框架推导出来的另一个模型也会拟合这组数据,这种可能性总是存在的。对假设检验的逻辑和含义的考量,与这里所提及的问题,直接相干。

到目前为止,建构之间的关系是我们的讨论所涉及的主题。在一些情形下,一项研究也可能会关注一个建构和一个可观察变量之间的关系。例如,我们可能想研究保守主义和年龄,或自尊与性别之间的关系。这种情形的一个特例是,在建构验证的语境下,我们研究一个建构和一个可观察变量之间的关系;我们把它称为"已知组别"法。克伦巴赫和米尔将这种方法称做"组别差异"(Cronbach & Meehl,1955),并描述道:"如果我们对一个建构的理解让我们期待:两个组别在测验上存在差异,那么,我们应对这个期待进行直接检验。"(Cronbach & Meehl,1955: 287)对该假设(即该期待)的支持,表示对各种推论的效度的支持,这些推论的基础是对该建构的测量的应答。

在建构验证的过程中使用已知组别法,特别容易让我们犯一

种错误,一种逻辑学家称做"肯定后项"的谬误(参见第9章)。在建构验证的背景下,当我们把已知组别在一个既定度量上的差异,看成是一个勘测当前建构的度量效度的证据时,我们就犯了"肯定后项"的谬误。为一个量表选择题器,或许是犯这种谬误的最臭名昭著的例子,因为在已知组别中,它们有所区别,然后,在这个量表的组别差异,又被当成是建构效度的证据。不幸的是,在社会行为研究中,我们经常会遇到这样的做法。例如,在许多性别角色量表中,选择题器的基础都是它们对男性和女性的鉴别能力。(有关综述,参见 Constantinople,1973)现在,认为一个既定量表测量男性气质,并假设,举例来说,在这个量表上,男性得分高于女性,这是一码事;认为一个量表测量男性气质,是因为男性得分高于女性,这完全是另一码事,而且是一种谬误。正如康斯坦丁诺普(Constantinople,1973)所指出的:

> 从概率上讲,大脚趾的长度可以分辨男性和女性;不过,一个女性比大多数女性的大脚趾长一点,这是否让她变得不够"女性气质"呢?并且,在一组具有类似的关键内容的题器上,由于她的得分偏差较大,所以她就变得不够"女性气质"。我们是否对这个结论具有更大的信心?

## 趋同及判别验证

一个研究者计划测量一个建构时,他面临的任务是,从大量令人眼花缭乱的不同测量方法和流派中作出选择。例如,当他准备测量一种态度时,他可能会诉诸自我报告、投射技术、可观察行为基础之上的推论、生理反应、社会计量等,从中选择一种来完成任务。[1]

在选定一种流派之后,他仍然要面对各种方法的选择问题。假定他决定使用自我报告法来测量态度,他究竟应该使用累加评分量表、等间距量表、古特曼量表、Q-分类问卷、清单表还是开放问卷?(对一些主要测量流派的描述,参见第6章)

与被访者、研究者以及研究背景有关的一系列因素,都可能对

---

① 对态度测量的不同流派,对使用多种流派的重要性的较好讨论,参见 Cook 和 Selltiz (Cook & Selltiz,1964)。

任何测量方法(或数据收集技术)的效度带来潜在的威胁。当他采用单一的方法来测量一个建构时,选项集、对研究者期望的反应、特定的研究背景等,是否(或在多大程度上)影响被访者的应答?就是一个无法确定的问题。[1]

研究者经常会使用单一的方法(例如,累加评分量表,参见第6章)来测量几个建构(例如,对黑人的态度、自我概念、保守主义),然后,计算这些分数之间的相关关系,以便研究这些建构之间的关系。在这些情形下,可观察的相关系数在很大程度上(甚至完全)取决于用于测量这些建构的特定方法。

依赖单一的测量方法会给我们带来风险,我们可以在群际刻板印象的领域中举一个例子。在这个领域中,存在相对较多的研究文献,其中,不同的研究者宣称,针对各种群体(例如,土耳其人、犹太人、黑人)的刻板印象,在不同被访者之间、不同的时点之间,他们发现了较高的一致性和稳定性。在对这个研究领域的一个评论中,埃利希(Howard J. Ehrlich)和莱恩哈特(James W. Rinehart)认为,许多研究的结论一致性,很大程度上源自这些研究都依赖"刻板印象清单表"(Ehrlich & Rinehart,1965)。为了支持自己的命题,埃利希(Ehrlich)和莱恩哈特(Rinehart)证明,是我们给被访者一张刻板印象清单表,还是给他们一份开放问卷而定,被访者的应答存在重要的差异。(Ehrlich & Rinehart,1965)例如,与使用开放问卷的被访者相比,使用清单表的被访者会将更多的特质分配给目标群体。而且,使用开放题格式的被访者所产生的特质清单和使用清单表格式的被访者所产生的特质清单之间,存在显著差异。

为了克服使用单一方法所带来的困难和偏差,许多研究者提倡采用多种方法来测量一个建构。一些研究者将这种流派称为"多重操作化"(例如,Garner et al. ,1956);另一些研究者称之为"三角测量法"(例如,Denzin,1978,第10章);还有一些研究者将它称为"多方法流派"(Campbell & Fiske,1959)。不管称谓如何,只有这样一种流派才能许下一个诺言:甄别出各种源自一个特定方法或一个方法与其他因素间相互作用而产生的偏差。

在一篇会议论文中,坎贝尔和菲斯克(Campbell & Fiske,1959)提出"趋同效度"和"判别效度"两个概念。趋同效度是指,旨在测

---

① 这些问题及相关的问题,将在第10和11章中予以讨论。

量同一个建构的不同方法(倾向于采用差异最大的方法)之间的一种趋同。例如,一个纸笔度量和一种投射技术,它们都设计为测量焦虑,如果它们之间存在高的相关关系,那么,这就构成趋同效度的证据。换句话说,趋同效度是指采用多个方法来证实一个建构的测量。

判别效度是指建构的独特性,由旨在测量不同建构的各种方法之间的趋异性来说明。例如,旨在测量两个不同建构(例如,焦虑和内倾)的两个量表之间的相关系数不能太高,否则的话,我们就会怀疑:它们是否测量了不同的建构?

人们常常"同名和异名谬误"的标签下讨论与建构的独特性或缺乏独特性有关的谬误。因为不同的事物被冠以相同的名称,然后就误以为它们是同一个事物,这样的信念被称为"同名谬误";反之,因为事物被冠以不同的名称,然后就误以为它们是不同的事物,这样的信念被称为"异名谬误"。在社会行为研究文献中,同名和(或)异名谬误的例子不胜枚举。(饶有兴趣的讨论和例子,请参阅 Kelley,1927: 62-64;Hartley,1967)

在测量建构的语境下,我们经常碰到的同名谬误的形式是,据称是测量同一个建构的各种量表之间,存在较低的相关系数;反之,我们经常碰到的异名谬误的形式是,据称是测量不同建构的各种量表之间,存在较高的相关系数。一个切题的个案是各种用于测量一个"知觉方式"的度量,称做"场依赖—独立"(FDI)。据称是测量 FDI 的各种度量之间的相关系数,变化的范围是中等负相关到中等正相关,中位数约为 0.4(参见 Arbuthnot,1972;Witkin et al.,1962: 44-45),这种情形让克伦巴赫评论道:"当假定是测量相同事物的测验之间,在一些群体中的相关系数是零,甚至是负值时,很明显,我们就不能依赖它们来从事任何研究。"(Cronbach,1970: 628;同时参见 Arbuthnot,1972)

简而言之,请注意,我们把 FDI 看成是一种人格建构,不同于能力建构。这样的话,FDI 的度量和能力度量之间的相关系数,应该不高;这才能为判别效度提供证据。前面曾提到过,由于缺乏趋同效度,我们不可能对 FDI 的度量和能力度量之间的相关系数,作出一个明明白白的表述。但已经证明,有些 FDI 度量和能力度量之间存在高度相关关系,弗农(Philip E. Vernon)对此评论道:"和大范围的空间测验之间存在高度正相关关系,几乎是一件令人难

堪的事情。”(Vernon,1972: 368)

让我们来看另一个例子。关于自尊的各种度量的趋同效度,存在93项研究,怀利(Ruth C. Wylie)对它们进行了综述。(Wylie,1974)在引用这项综述时,布里格斯(Stephen R. Briggs)和奇克(Jonathan M. Cheek)说道:“设计来评估总自尊的各种量表之间的相关系数,变化范围为0~0.8,相关系数的均值只有0.4。”(Briggs & Cheek,1986: 131)毫不奇怪,他们的结论是:“在人格心理学领域,自尊测量的研究状况已经变成一处伤疼。”(Briggs & Cheek,1986: 131)

在讨论趋同效度和判别效度时,“高”相关系数和“低”相关系数常常是参照系,但是,这些术语的含义是什么,却没有明确说明(前面我们曾这么做过)。自不必说,怎样才算是高相关系数,怎样才算是低相关系数?缺失这样的准则,就会打开了一扇门,它通向歧义,通向研究者之间的互不赞同,甚至通向单个研究者的著作之间的前后不一。后者的一个示例来自一项研究,它考察上级、同事和自己对岗位绩效的评估。据报告,对岗位绩效的三个方面,上级评估和同事评估之间的相关系数分别是0.52、0.53和0.65,劳勒(Edward E. Lawler Ⅲ)认为,它们表明该研究具有“较好的趋同效度”(Lawler,1967: 374)。然而,在报告看似是同一个研究的另一个方面时,劳勒指出,对岗位绩效的两个方面,上级的评估之间的相关系数是0.56,这次,一番阐释之后,他写道:“这个相关系数不太高,这表明,经理们无法辨别绩效因子和能力因子。”(Lawler,1966: 158)

## 多特质多方法矩阵

为了研究度量的趋同效度和判别效度,坎贝尔和菲斯克(Campbell & Fiske,1959)提出应分析“多特质多方法”(MTMM)矩阵。MTMM矩阵是通过两个或多个不同方法进行测量、两个或两个以上特质之间的相关系数矩阵。图4.6显示的是三种特质、三种方法的矩阵。

我们只在该图的一部分指明了相关系数,空出的部分用来标识MTMM矩阵的不同单元。举例来说,令$A$表示焦虑,$B$表示羞怯,$C$表示快乐;令方法1表示纸笔问卷,方法2表示一种投射技术,方法3表示一个临床心理学家的评估。这样,$A_1$表示用纸笔问

卷来测量的焦虑，$B_1$ 表示用纸笔问卷来测量的羞怯，以此类推，$C_3$ 表示用临床心理学家的评估来测量的快乐。①

沿着主对角线（从左上方到右下方）的相关系数是每一个度量的信度系数（例如，$r_{A_1A_1}$ 是利用纸笔问卷所测量的焦虑的信度系数；参见第 5 章有关信度的讨论）。

围在实线三角形里面的相关系数，是同一方法测量的不同特质之间的相关系数，因而是"异特质—同方法（HTMM）"三角。例如，$r_{B_1A_1}$ 是用纸笔问卷测量的羞怯与焦虑之间的相关系数。

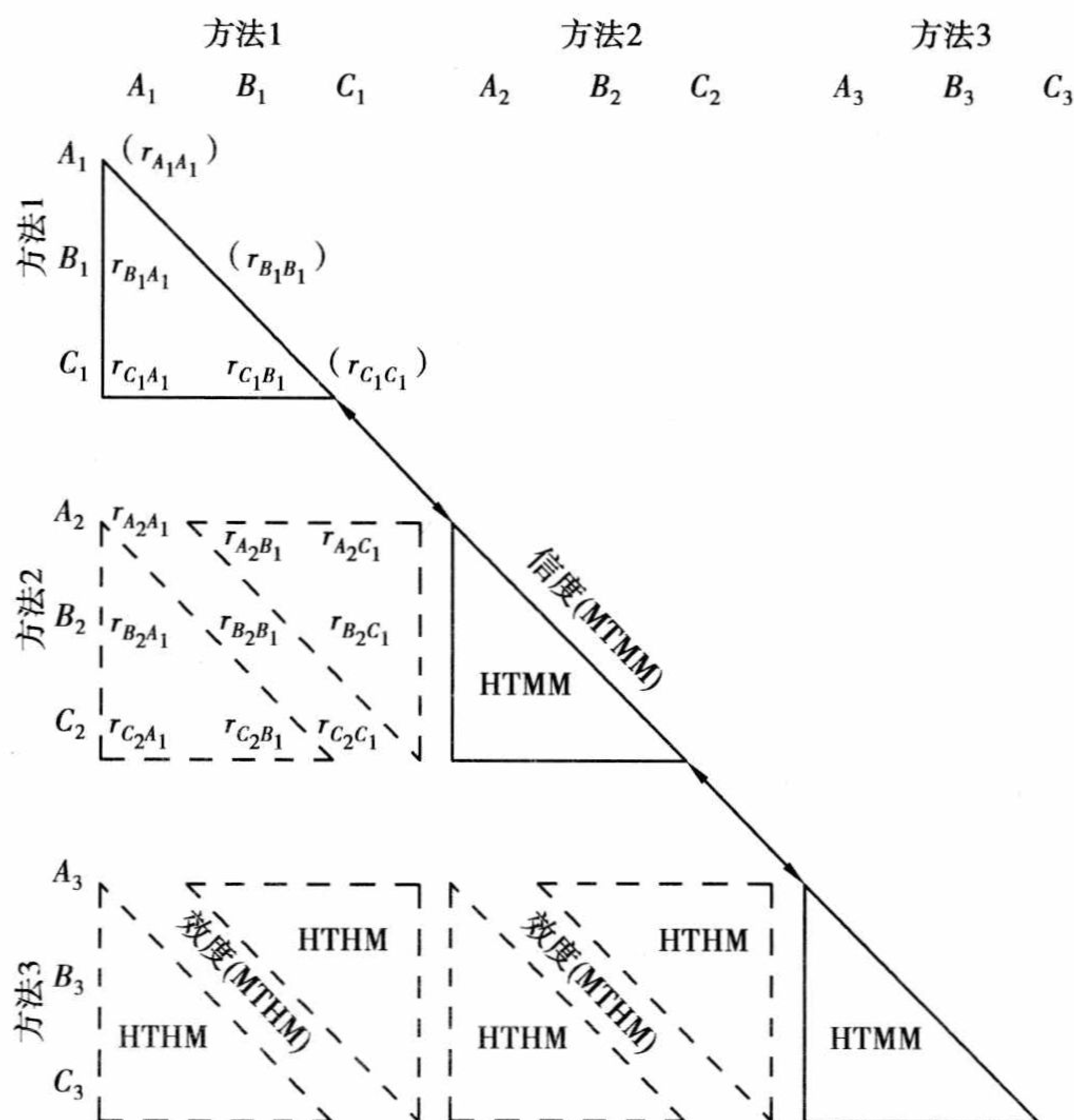



图 4.6

围在虚线三角形里面的相关系数，是不同方法测量的不同特质之间的相关系数，因而是"异特质—异方法（HTHM）"三角。例如，$r_{C_2A_1}$ 是由投射技术测量的快乐与用纸笔问卷测量的焦虑之间的相关系数。

① 什么才构成不同的方法？这个问题将在下面予以讨论。

两个 HTHM 三角形之间的对角线上的相关系数,是由不同方法测量的相同特质之间的相关系数,因而是“同特质—异方法(MTHM)”三角。例如,$r_{C_2C_1}$是由投射技术测量的快乐与由纸笔问卷测量的快乐之间的相关系数。MTHM 对角线被称做“效度对角线”。

我们可以从 MTMM 矩阵中收集一些信息,现在,我们将指明它们的一些方面。例如,对照 $r_{B_1A_1}$和 $r_{B_2A_1}$。在这两个情形下,我们都在处置羞怯和焦虑之间的相关系数。但 $r_{B_1A_1}$表示由同一方法(例如,方法 1)测量的不同特质之间的相关系数,而 $r_{B_2A_1}$表示由两种不同方法(例如,$B$ 由方法 2 测量,$A$ 由方法 1 测量)测量的羞怯和焦虑之间的相关系数。如果 $r_{B_1A_1}$大于 $r_{B_2A_1}$,可得出结论:这是源自一个共同方法(方法 1)的效应,它同时测量 $B$ 和 $A$,并计算出相关系数。

现在让我们看一下 MTHM 对角线(即效度对角线)。因为它们表示用不同方法测量的同一个建构之间的相关系数,这些相关系数应当比较高,这样才能表示存在趋同效度。现在,让我们把 MTHM 对角线上的一个相关系数和 HTHM 三角形中对应行或列上的另一个相关系数进行比较,如对比 $r_{B_2A_1}$ 和 $r_{A_2B_2}$,或者对比 $r_{B_2A_1}$ 和 $r_{A_2A_1}$。请注意,在所有例子中,我们使用的是相同的两种方法(即方法 1 和方法 2)。但是,$r_{B_2A_1}$表示用两种方法测量同一个特质的相关系数,其他相关系数则是用相同的两种方法来测量不同的特质时所得的相关系数。这样,我们将期待,$r_{A_2A_1}$大于对应行或列上的其他相关系数(例如,$r_{B_2A_1}$),否则的话,所涉及的度量的效度就值得怀疑。

沿着上一个段落中所指明的路线进行推理,坎贝尔和菲斯克(Campbell & Fiske,1959)获得了一组指南和准则,用于研究 MTMM 矩阵中的趋同效度和判别效度。但是,有批评认为,他们的指南和准则有局限、很含糊,而且是建立在有疑问的假定之上。(参见 Althauser & Heberlein,1970;Jackson,1969;对 Althauser & Heberlein 的评注,参见 Alwin,1974;回应,参见 Althauser,1974)有人还提出了分析 MTMM 矩阵的其他方法。(综述,参见 Alwin,1974; Schmitt et al.,1977;Schmitt & Stults,1986)目前的共识似乎是,证实性因子分析(CFA)是信息量最大的分析方法。将 CFA 应用于 MTMM 矩阵,其他除外,我们还可以研究特质、方法以及它们之间的相关系数的效应。在第 23 章中,我们会举例说明把 CFA 应用到 MTMM 矩阵之

上，并提供应用这种方法进行研究的参考文献。

*什么是不同的方法？*

如前所述，坎贝尔和菲斯克（Campbell & Fiske，1959）把趋同效度界定为测量同一个特质的、最大不同的方法之间的高相关系数。但是，什么才是不同的方法？这个问题并不存在明晰的准则，更不用说最大不同的方法了。这个问题的一部分在于，在一个既定研究中，由于特定的特质、背景、被访者等原因，看似不同的方法可能是相互相关的。例如，在一个特定背景中测量特定的特质的不同方法，可能会受到光环效应、社会赞许性等效应的影响，因而，和表面上看起来的差异相比，它们之间的相互差异可能要小一些。（关于这个论点的讨论，参见 Jackson，1969）

在认定什么才是 MTMM 矩阵中的不同方法时，不同的研究者似乎采用了不同的准则。对 MTMM 矩阵的分析和阐释而言，这具有深远的意义，因此，我们将举几个内容相当不同的例子，说明什么是不同的方法。下面是对阿维森（William R. Avison，1978：441）所分析的 MTMM 矩阵的一个描述：

> 这三个特质分别是被访者在家庭（作为孩子）、学校（作为学生）和工作（作为员工）中对决策过程的参与感。每个特质由三种方法来测量：自认的参与自由、自认的参与影响和参与的实际频次。

请注意，阿维森（Avison，1978：441）自己承认"在一定程度上，该数据不同于 MTMM 情境下的更标准的范例"。不过，有关参与决策的知觉问题，有关自认参与的自认影响问题，把它们看做是不同方法，最低限度来说，也是值得怀疑的。

关于"什么是不同的方法"的令人质疑的概念，来自瓦诺斯和劳勒（Wanous & Lawler，1972）的一项研究，他们采用 MTMM 矩阵，研究工作满意度的测量和意义；这是另一个例子。扼要地说，它们要求被访者在 23 个题器上、分 5 次来评估自己的工作，每一次评估采用不同的参照系（例如，"每一个品质或特征多大程度上体现在您的工作中？""您认为，每一个品质或特征应在多大程度上和您的工作相关联？"（参阅 Wanous & Lawler，1972：98）使用这些评分，他们推导出若干指数（例如，对工作的现状评估与规范评估之间的差

距),并在一个 MTMM 矩阵中使用下列四个"方法":(a)满意度;(b)现状;(c)理想减现状;(d)规范减现状。[①]

应当注意到,瓦诺斯和劳勒将操作性定义等同于方法。(具体例子,参见 Wanous & Lawler,1972: 98, 102)[②]卡勒伯格(Arne L. Kalleberg)和克鲁格(James R. Kluegel)对瓦诺斯和劳勒的数据重新进行了分析,他们也重复了这种做法,有下面的表述为证:"我们用这四种方法或操作性定义来测量工作满意度的四个特质或剖面。"(Kalleberg & Kluegel,1975: 6)尽管如此,卡勒伯格和克鲁格(Kalleberg & Kluegel,1975: 1)还认为,瓦诺斯和劳勒从他们的数据中得出"错误的推论"。

使用 CFA,卡勒伯格和克鲁格(Kalleberg & Kluegel,1975: 7)发现,瓦诺斯和劳勒所使用的四个"方法"之间存在"高相关关系;事实上,第三个和第四个测量方法之间的相关系数高达 0.96,我们最好把它们看做是一种测量方法"。如果大家还记得瓦诺斯和劳勒究竟把什么看做是不同的测量方法的话,这样的高相关系数,也就不足为奇了。例如,上述 0.96 是两个指数之间的相关系数,其中一个指数是建立在工作"理想"和"现状"之间的差距上,另一个指数是建立在"规范"和"现状"的差距之上。

在态度研究中,我们经常用到 MTMM 矩阵。例如,奥斯特姆(Ostrom,1969)使用四种方法(例如,累加评分量表、等间距量表)来测量人们对教会态度的情感、行为和认知成分。巴戈兹(Richard P. Bagozzi)采用 CFA,重新分析了奥斯特姆的数据(Bagozzi,1980a: 136-146);其他除外,他发现有两个度量之间的相关系数"几乎为 1"(Bagozzi,1980a: 142),并且,"我们可能不会拒绝这个假设:这两种方法等价"(Bagozzi,1980a: 144)。

如果不深入实质问题和测量问题,我们就不可能解释这样的研究结果。但请注意,在这里,我们的目的仅限于讨论不同的研究者如何回答这个问题:什么才是不同的方法? 因此,指出一点就足以说明一切。尽管奥斯特姆采用四种不同方法来测量对教会的态度(Ostrom,1969),但它们拥有共同的成分(例如,自我报告、纸笔度量、对态度表述指出赞同—不赞同的形式)。

---

① 应用分数差的有关问题,将在第 13 和 21 章中加以讨论。

② 操作性定义将在第 8 章中加以讨论。

在 MTMM 矩阵中，我们经常把不同来源的评估（例如，自我评估、同事评估、上级评估）看做是不同的方法（Fiske，1949；Schneider，1970）。卡瓦纳等人（Kavanagh et al，1971：36）“质疑把自我评估用在多特质—多方法分析中的价值”，理由是“一个特定的行为可能具有不同的意义，对当事人而言是动作意义，对观察者而言是行动意义”。我们从上面得到的结论是什么？涉及被评估事物的不同方面的自我评估，高于上级评估吗？因而我们不应当把它们看做是不同的方法吗？另一方面，例如，自我评估和同事评估是相同的测量方法吗？对这些问题，并不存在简单的答案。实际上，在什么是不同的测量方法，在为解决这个问题寻找准则的过程中，我们已经兜了一个圈。沃林斯（Wolins，1982：63）一直致力于解决“什么是不同的方法”这个问题，他评论道：“各种特质之间相关，这个观点说得通，但说各种方法之间相关，对这个观点，我就无法从认知上加以领会了。”

总之，在 MTMM 矩阵的使用、分析和阐释中，存在内在的复杂性。我们希望上述讨论有助于凸显这种复杂性。应用 MTMM 矩阵来评估测量，在对这种做法进行评注时，坎贝尔观察到，这是一个“让人丢脸的经历，带给人谦卑和谨慎”（Campbell，1960：552）。如前所述，对 MTMM 矩阵的分析将在第 23 章中加以说明。

## 内容效度的一点注释

在第 3 章介绍效度概念时，我们曾经指出，分别讨论验证过程的不同方面，虽然比较方便，但有可能带来过度简化和相互混淆的危险。特别是，它可能会让我们形成一个根深蒂固的错误观念，以为存在不同“类型”的效度。在讨论内容效度，并把它和建构效度作区分时，这种混淆变得尤其明显。造成混淆的原因在于，内容效度根本不是一种效度。

为了澄清当前的状况，我们必须指出，到目前为止，对关注成就测量的教育心理学家和教育家而言，内容效度几乎成了他们的唯一领地。当他们关注成就的一个度量时，他们事实上期待内容效度的证明；当他们关注特质、属性等的度量时，他们期待建构效度的证明。论述测量的教材，在很大程度上强化了这种观念。例如，桑代克和哈根（Thorndike & Hagen，1977：58）认为：“我们应当

清楚,主要对成就的度量而言,合理效度或内容效度才是重要的。”

近年来,有些机构和专业工作者专注于遴选员工和平等就业机会,在他们的工作中,有关内容效度的概念开始扮演重要的角色。有关在招工、认证等工作中使用考试的法令,以及由美国联邦机构颁布的、遴选员工的指南,都极大刺激了人们对内容效度的兴趣。例如,美国有关部门所采用的《遴选员工统一指南》,就对内容效度的证明作了规定。除了其他之外,它指出:

> 为了证明一个遴选程序的内容效度,一个用户应当说明:出现在该遴选程序中的行为是当前岗位行为的一个代表性样本;或者,该遴选程序提供了该岗位工作产品的一个代表性样本。(the Equal Employment Opportunity Commission et al,1978:38302)

诸如上面的各种表述,无论是出现在成就测验,还是出现在遴选程序的情境下,最重要的是要注意到,它们与效度的定义不一致。正如我们在第3章的开篇所讨论的,效度是指就有关分数所作出的推论,而不是只对一个测量工具的内容进行评估。这并不是说,一个测量工具的内容不相干。很清楚,它是至关重要的,但它并不构成这个测量工具的效度证据。

有些人公开关注标为“内容效度”的事物,除他们之外,大多数社会行为学家对他们想要测量的内容领域,流露出一种满不在乎的态度,这种情况仍然存在(参见 Bohrnstedt,1983:98)。

在阅读有关内容效度的材料时,我们会形成一个印象。和这种印象相反,内容相干性并不局限于测量成就。在缺乏内容领域时,我们如何考量一个建构?如何进而开发这个建构的一个度量?一个建构的定义本身就暗含一个内容领域。例如,想一下“保守主义”这个建构。我们关注的是“气质保守主义”“情境保守主义”“政治保守主义”,还是“作为一个哲学家的保守主义”(参见 Rossiter,1968)?自不用说,保守主义的一个特定度量的内容,取决于我们想测量的保守主义究竟是前面的哪一个概念,或者是其他概念。在本章的前面章节(特别是“逻辑分析”一节),我们详细讨论了一个要求:一个度量的内容须与该建构的定义一致。这里我们就不再赘述了。

正是上述考量,促使许多研究者反对“内容效度”的概念。(参

见 Fitzpatrick, 1983; Guion, 1977, 1978; Messick, 1975, 1980, 1981; Tenopyr, 1977)而且,这些研究者和其他一些研究者强调,我们必须从自己想要测量的那个建构的视角,来评估所有度量。因此,梅西克(Messick, 1975:957)主张"所有测量都应以建构为参照"。泰诺佩尔(Tenopyr, 1977:48)则争辩说:"和预测关联的任何推论和测验分数关联的所有推论,都必须建立在它们背后的建构之上。"盖恩(Guion, 1978:209)宣称:"有关分数的备择阐释,我们形成了建构效度的一个原则。在评估分数时……我们必须按照这个原则来评估赋值给绩效的分数。"

作为对上面所引用的这些论证的回应,近来修订的《教育与心理测验标准》(American Psychological Association, 1985: 11)写道:

> 因此,分类在"内容相关"类别下的各种方法,应当经常参考该测验背后的心理建构,参考该测验的内容特征。测验内容与测验建构之间,常常并没有分明的界限。

## 结束语

在本章以及上一章,我们对验证过程说了很多,见木不见林的危险,已经清晰可见。因此,十分重要的是,我们觉得应强调一下,我们想要传达的要点是,验证过程是科学研究不可或缺的一部分。下列问题随之而来。

理论、研究设计和分析对验证过程具有直接的影响,正如验证过程对它们也具有直接影响一样。因此,如果不参考或不了解研究项目的其他方面,那么各种度量的设计注定会失败。正是由于这个原因,我们才不断推荐大家参考本书的其他章节,同时,我们也十分清醒地意识到一种危险:这种老调重弹可能会引起大家的厌烦。

验证是一种复杂、持续不断的探索,它要求我们严肃和坚持。有一本书关注当前社会测量中所存在的各种问题,在它的前言中,主编们指出"可用于评估度量的各种模型的质量,远远超出这些度量本身的质量"(Bohrnstedt & Borgatta, 1981: 14)。博加塔(Edgar F. Borgatta)和博恩斯泰特(George W. Bohrnstedt)的意思并不是说,测量模型不重要,而是说,考虑到测量在社会行为科学中的现

状,对研究者来说,更重要的事情可能是"以更多的谨慎和深思熟虑,来推进度量的开发。开发好的度量,可能要花费数年的时间,而不是数日或数月"(Bohrnstedt & Borgatta,1981: 14)。

在社会行为研究中,鉴于我们使用和建构度量的随便方式,我们认为,恰当的做法是以克兰德尔(Rick Crandall,1973:52)的申诉作为结语:"随便生成新的量表,就是不负专业责任。"

# 5

# 信　度

我们将在本章讨论“测量精度”这个涉及面很广的话题。首先，我们将对信度在测量和研究中的状况作一个概括性的扫描，然后，我们将转向考察测量误差的不同类型和来源，看看它们如何与“信度”的概念与定义联系在一起。之后，我们将把古典测量理论作为一个示例来讲解，它是飞速发展的信度理论的一个例子，也是介绍信度基本概念的一个工具。紧跟其后，我们将综述信度估计最常见的一些方法，并结合手算或计算机程序来分析一些数据示例；在上述背景下，我们将强调内部一致的估值。随后，我们将综述在选择一个信度估值时所需要考量的各种因素。最后，我们将选择一些相关话题作为本章的结尾，包括信度标准、信度和效度之间的关系，统计、估计缺乏信度所带来的效应等。

## 信度在测量和研究中的状况

在社会行为科学中，和效度相比，我们把更多的注意力投在信度上。看起来是，很多人都把信度看做是首要的测量问题。这是一个错误的观念，它带来的有害后果可能会造成很大的影响。

和效度相比，信度更易于用数学公式进行表达，并且，我们比较容易取得较高的信度系数，因此，许多研究者和公众都会忽略一个事实，即信度是效度的一个必要、而非充分条件。也就是说，如果一个度量没有信度，它就一定没有效度；如果它具有信度，就作者或公众的使用目的而言，它不一定具有效度。较高的信度系数非常具有诱惑力，但也是一种危险，它会让我们错误地以为这是效度的证据。这可能是罗兹布姆（Rozeboom，1966：375）把信度称为

"贫者的效度系数"或"快餐效度"的原因。

这段话并不表示信度不重要,或者我们不必关注各种度量的信度,它们或者是我们自己正在使用的,或者是别人已经使用过的。恰恰相反,正如本章反复强调的那样,各种度量的信度极其重要,而且在任何研究题器中,它们都是不可分割的一部分。前面这些评述的目的是,我们要把信度放在一个恰当的视野下。

## 系统误差和非系统误差

最一般地来说,"信度是指测试分数远离测量误差的程度"(American Psychology Association,1985:19)。广而言之,在测量过程中会出现两种误差:系统误差和非系统误差。正如这些标签所暗示的,系统误差就是在重复测量中反复出现的误差,非系统误差(或随机误差)是指重复测量中不断变化、无法预测的误差。

例如,假定我们用一把尺子测量一个桌子的长度,为了评估精度,我们进行重复测量,看看不同测量之间的一致性。在这样的重复测量中,一致性会出现,这是我们的预期,但一定的变异度还是不可避免的。一个可能的情形是,我们没有把尺子每次都放在完全相同的位置上,或者,各次测量的读数会有所不同。这些误差越随机,它们对重复测量的影响就越不一致,越不可预测,它们和信度具有反向的关系。

在测量桌子的长度时,也有可能会出现恒定误差或系统性误差。例如,我们可能无意中用了一把尺子,它的刻度起点是1英寸而不是0英寸。很明显,不管测量重复多少次,这样的误差也不会发生变化。因此,就系统误差而言,重复测量的结果是一致的,这有助于测量信度的提高。但是,我们必须认识到,系统误差对效度有负面效应,这一点很重要。对社会行为测量的系统误差来源和非系统误差来源的详细讲解,可以参阅下列文献:Ghiselli,1981:242-247;Nunnally,1978:225-229;Stanley,1971;Thorndike,1951。

## 信度概念

三十多年前,如果有一个研究者想估计一个度量的信度,他会求教于一本学术著作或寻找如何做的指南;在特赖恩(Robert C. Tryon)看来,这位研究者将"面对一大批不同的公式,不知道如何进行下去。一个客观的行为度量在多大程度上能够稳定地区分不同

的个体？我们发现，在心理测试发展了50年之后，这个问题仍旧是扑朔迷离的”(Tryon，1957：229)。

就众多的新公式(包括特赖恩的公式)而言，信度估计的当前状况变得更混乱，更复杂。这种复杂性源自信度的不同理论表述，它的背后是有关真值和误差等问题的不同假定。正如下文将提到的，不同的信度估计关注不同的误差来源。因此，在一种流派的视野下，我们可以把它们看做是随机误差；在另一种流派的视野下，我们则把它们看做是系统误差，甚至忽略不计。所以，认为“一个度量只有唯一信度”的说法，并不恰当，潜在地会误导他人。相反，我们必须把信度和信度系数看做是广义的术语(American Psychological Association，1985：19)。

## “重复测量”概念的问题

对一个对象进行重复测量，这让物理学家能够评估一个度量的精度。在社会行为科学中，情形就不是这么简单了。在这里，我们所感兴趣的大多数变量，并不适用于进行重复测量，这样做没有意义，更不用提进行重复测量时所遭遇的各种实际问题。

测量这个动作本身就已经改变了待测的被访者，只是程度不等而已。比方说，一个人回答一个成绩测试中的问题，这已经构成一种学习和练习，并有可能带来其他变化，然后再影响到他对一个度量所试图测量的内容的应答。再举一例，一个态度调查在两个时点上实施，有一个人两次答案都一样，这可能不是(至少部分)因为他前后态度一致，而是因为他记得第一次调查时所给的答案。简而言之，“再测量(remeasurement)”几乎就是一个错误的命名，更不用说“重复测量(repeated measurements)”了。

由于各种各样的难题，在社会行为科学中，我们通常考察不同个体在一个群体中保持相对位置的程度来估计测量信度。一般来说，这是一个合理的方法，因为我们的兴趣常常是个体之间的差异(例如，个体之间在特殊特质或特征上的差异)，而不是个体内部的差异(例如，在几个不同时间点或不同情境下，一个个体在一个特殊特质上的差异)。

当我们需要就个体之间的差异形成比较性命题或者依据它们在一些度量上的分数进行决策时，我们就必须了解到：对这些分数，我们能够报以多大的信心，这是一个基本要求。一般来说，一

个度量的信度越高,我们对这个度量所得的分数的信心就越大(参见本章后面的“测量标准误”一节)。

## 古典测量理论

各种各样的信度理论都曾经有人表述过。我们的关注点是信度在研究背景下的作用,因此,我们不必综述这些理论,只关注信度估计的程序;另外,我们也关注信度对效度和统计分析的含义。不过,我们相信,举例说明一个信度理论如何得到表述,将有助于大家理解信度的理论概念,以及由此推导出的估计程序。由于历史的原因,并且对大多数其他理论而言,古典测量理论都是出发点,因此,我们将把它作为我们的示例。

自从斯皮尔曼提出“真值模型”(后来被大家称为“古典测量理论”)以来,它一直是占主导地位的理论,并指导着信度的估计。(Spearman,1904)这个模型经历了变迁和扩展,今天几乎已经没有人接纳它最初的形式了。为了介绍一些基本概念,介绍由它们如何导出信度的定义,我们会讲解古典测量理论的基本要素和假定。(古典测量理论的更详细讲解,参见 Gulliksen,1950;Lord & Novick,1968)若有需要的话,我们也会参考一些推广和另类概念,它们是用来处理古典测量模型的内在难题或问题的。

根据真值模型,我们认为,一个观测值由两个成分组成——真值成分和误差成分。用符号表示如下:

$$X = T + E \tag{5.1}$$

其中,$X$ 是带有误差的观测值,$T$ 是真值,$E$ 是随机误差。

从概念上来说,我们可以把真值看做是在理想的(或完美的)测量条件下得到的分数。由于这样的条件从不存在,因此观测值总是包含着一定量的误差。

虽然公式(5.1)具有理论意义,但我们不能用它来估计一个观测值中的误差量,反过来,我们也不能用它来估值这种分数的精度。原因在于:公式(5.1)包含两个未知数(即 $T$ 和 $E$),如果不作进一步的假定,我们就不可能求解。在古典测量理论中,我们假定所测量的特质恒定,且测量误差是随机的。这样的话,如果我们把一个人测量几次,假定他(或她)暂时没有发生变化,那么我们就可

以得到一组类似公式(5.1)的方程,每个方程都由一个相同的真值(因为我们已经假定它恒定)构成,但观测到的分数变化则是由于误差变化而造成的。由于是随机的,多次重复测量的测量误差的均值期望值为0,即:

$$E(E)=0 \tag{5.2}$$

其中,$E(\ )$是期望值的运算符号,是指随机变量$E$的期望值。一个随机变量的期望值是它在一个无穷大的重复随机样本上的长程均值(有关期望值以及它在统计学和概率论中位置的讨论,参见Edwards,1964;Hays,1988:164-166,866-872)。可见,真值等于观测变量的期望值,即无穷多的重复测量所得到的观测值的均值。

$$T=E(X) \tag{5.3}$$

这个概念的基础是一个站不住脚的假定:当我们重复测量一个个体时,他的真值保持不变。因此,如前所述,我们采用了另一种不同的方法来估计信度。我们不是重复测量同一个个体,而是采用相同的量表(或平行格式)一次(或两次)测量很多个体,然后用所获得的信息来估计该量表的信度。

在不失普遍性的前提下,在下面的讲解中,我们将采用离差值而不是原始分。这样,公式(5.1)可以改写为:

$$x=t+e \tag{5.1'}$$

其中,$x=X-\overline{X}$,$t=T-\overline{T}$,$e=E-\overline{E}$,即原始分减去各自的均值。

一般地说,观测值上的个体差异可能来自他们之间在所测量的特质上的真实差异,也可能来自各种误差(例如,猜题、疏忽、粗心等)。这样看来的话,我们的目标就变成:把观测值的方差分解成真实方差和误差两个部分。在讲解如何做到这一点之前,我们有必要首先考察真值与误差之间的关系。请注意,我们假定误差是随机的,由此推知,真值和误差之间的相关系数的期望值为0。还可以推知的是,当我们对多个个体的分数求均值后,误差相互抵消(误差的均值为0)。可见,观测值的均值等于真值的均值,即:

$$E(X)=E(T) \tag{5.4}$$

请注意,一个观测值是真值和误差的复合分,因此,我们可以

把观测值的方差表示为真值和误差和的方差①：

$$\begin{aligned}\sigma_x^2 &= \sigma_{(t+e)}^2 \\ &= \sigma_t^2 + 2\sigma_{te} + \sigma_e^2 \end{aligned} \tag{5.5}$$

其中，$\sigma_t^2$ 为真值方差；$\sigma_e^2$ 为误差方差；$\sigma_{te}$为真值与误差的协方差。由于真值与误差之间的相关系数(因而，协方差)②为0(参见前面章节)，所以观测值方差等于真值方差与误差方差之和，即：

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2 \tag{5.6}$$

现在，让我们看看观测值(即 $t+e$)与真值($t$)之间相关系数的一些属性③：

$$\begin{aligned}r_{xt} &= \frac{\sum (t+e)t}{N\sigma_x\sigma_t} \\ &= \frac{\sum t^2 + \sum te}{N\sigma_x\sigma_t} \\ &= \frac{\sigma_t^2 + \sigma_{te}}{\sigma_t\sigma_x}\end{aligned}$$

[因为 $\sigma_{te}=0$]

$$= \frac{\sigma_t^2}{\sigma_t\sigma_x} = \frac{\sigma_t}{\sigma_x} \tag{5.7}$$

以语言表述就是：观测值与真值之间的相关系数等于真值的标准差和观测值的标准差之比。

现在，相关系数($r$)的平方是指和它相关的另一个变量所共享的方差的比例，或者说，一个变量的方差中可以由另一个变量解释掉的比例(参见第17章)。在当前的语境下，相关系数的平方是指观测值的方差由被测量的被访者的真值差异所解释的比例，即：

---

① 我们将在本章后面的"$\alpha$ 系数"一节讨论一个复合分的方差。

② 有关方差、协方差和相关系数的讨论，参见第17章。

③ 在信度理论的框架下，在讲解和推导各种相关系数方程时，有些学者(例如，Ghiselli et al., 1981; Gulliksen, 1950)使用 $r$(样本统计量)，有些学者(例如，Allen & Yen, 1979; Lord & Novick, 1968; Zeller & Cannines, 1980)使用 $\rho$(参数)。为了和后面章节衔接，我们使用 $r$。

$$r_{xt}^2 = r_{xx} = \frac{\sigma_t^2}{\sigma_x^2} \tag{5.8}$$

其中,$r_{xx}$是度量 $X$ 的信度。那么,下面就是一个度量的信度定义:它是真值方差与观测值方差之比。

由信度系数公式(5.8)的定义可推知,它的平方根等于观测值与真值的相关系数(公式 5.7),也称为“信度指数”。由于信度指数是指一个观测值与一个潜变量(建构)的取值之间的关系,它也称做一个测量的“理论效度”(Lord & Novick,1968: 261),或者“认知相关系数”(Northrop,1947:第 7 章)。

使用公式(5.6),真值方差可以表示为:$\sigma_t^2 = \sigma_x^2 - \sigma_e^2$。带入前面的公式(5.8)的分子中,我们就可以得到信度的另一个表达式:

$$\begin{aligned} r_{xt}^2 = r_{xx} &= \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} \\ &= 1 - \frac{\sigma_e^2}{\sigma_x^2} \end{aligned} \tag{5.9}$$

由此可见,信度的取值范围是 0 到 1。当所有观测到的方差均来自真值时,信度为 1,即测量不存在随机误差。另一极端是,当所有观测到的方差均来自随机误差时,信度为 0。

请注意,这个模型并不区分真值方差与系统误差方差(例如,应答方式源自所用的特定测量方法的方差)(对后者,参见第 4 章“趋同效度和判别效度”一节)。换句话说,信度(按上述定义)实际上是系统性方差占观测到的方差的比例。还需要注意的是,信度($r_{xx}$)的定义是一个相关系数的平方(即观测值与真值的相关系数的平方)。因此,我们把信度系数(而不是它的平方)阐释为是系统性方差占观测值方差的比例。例如,$r_{xx} = 0.8$ 意味着观测值方差中 0.8(或 80%)是系统性方差;$1 - 0.8 = 0.2$ 是源自随机误差的方差比例。

如前所述,在估计信度时,由于我们可能关注不同的误差来源,使用广义的“信度”术语,应该更合适一些。现在,我们已经证明,信度估值本质上是相关系数的平方,如果我们不指明用来进行信度估计的样本所对应的总体的话,那么,我们就应当避免说:一个量表具有唯一信度。

正如我们在第 3 章中所指出的那样,相关系数是总体有别的。由此推知,随着研究总体的变异度的高低不同,相同测量量表的信

度也会发生变化。有些学者完全讨厌报告他们所经常(或不经常)使用的量表的信度估值,他们只报告量表手册中的(或其他学者的)信度估值。相比较而言,那些信息或许是有用的,但我们必须认识到,相干的信度估值只能是来自当前研究所用样本的估值。只有这个信度系数,才能用于计算其他统计量(例如,测量的标准误,参见下文);只有这个信度系数,才能用于解释一些结果(例如,所研究的变量之间低于预期的相关系数,参见下文衰减"校正"一节)。

### "平行度量"的概念

尽管上述方程具有令人感兴趣的属性,但我们仍不能用它们来估计信度,因为它们包含一个无法观测的元素——真值方差。一个可行的解是,利用所研究的同一个属性的两个度量之间的相关系数,作为它们当中任何一个度量的信度估值。很明显,这两个度量应当相似,在古典测量理论的语境下,这意味着它们应当平行。

### 平行度量

两个度量($X_1$ 和 $X_2$)平行,当

$$X_1 = T + E_1$$
$$X_2 = T + E_2$$
$$\sigma^2_{e_1} = \sigma^2_{e_2} \tag{5.10}$$

这就是说,当两个度量具有相同的真值,且具有相等的误差方差时,它们就是平行的。(Novick,1966)因此,这两个度量的均值与方差也相等。请注意,我们已经假定误差是随机的,可推知,和平行度量关联的误差之间不相关,它们与真值也不相关,无论是相同度量的真值还是一个平行度量的均值,即:

$$r_{e_1e_2} = 0 \tag{5.11}$$

$$r_{e_1t_1} = r_{e_1t_2} = r_{e_2t_2} = r_{e_2t_1} = 0 \tag{5.12}$$

利用上面的等式,可见,两个平行度量之间的相关系数,就是它们当中任一个度量的信度估值。在每一个形式中,我们把观测值表示为真值和误差的复合分,两个平行形式之间的相关系数是:

$$r_{x_1x_2} = \frac{\sum (t + e_1)(t + e_2)}{N\sigma_{x_1}\sigma_{x_2}} = \frac{\sum t^2 + \sum te_1 + \sum te_2 + \sum e_1e_2}{N\sigma_{x_1}\sigma_{x_2}} = \frac{\sigma_t^2 + \sigma_{te_1} + \sigma_{te_2} + \sigma_{e_1e_2}}{\sigma_{x_1}\sigma_{x_2}} \quad (5.13)$$

[因分子的后三项为 0,且 $\sigma_{x_1} = \sigma_{x_2}$]

$$r_{x_1x_2} = \frac{\sigma_t^2}{\sigma_x^2}$$

它与前面给出的信度定义一致。

如果采用平行度量模型的话,针对所研究的属性,我们需要管理两个平行度量,并用它们之间的相关系数来估计其中任一个的信度。然而,这意味着,有关平行度量的各种严格假定需要得到满足,这种情形十分罕见。出于种种原因,我们发现,原初表述的真值模型十分罕见,且具有很大的局限性。

## 真值模型的变种

已经出现的各种不同的表述,我们可以把它们看成是真值模型的变种。它们共享一个基本目标,即建构各种旨在“测量相同现象”的度量。这些模型的不同之处在于,它们所放宽的平行度量模型的假定不同。在这个阶段,我们只能就这类模型的一些作一个简要的刻画。在第 23 章(测量模型)中,我们将应用证实性因子分析,说明如何才能确定这些模型是否与数据拟合。

真值等价度量。当两个度量具有相同的真值时,我们把它们称为“真值等价”。尽管这是和平行度量模型共享的一个假定,但真值等价度量没有假定误差方差相等(Novick & Lewis,1967)。

准真值等价度量。进一步放宽假定,除了误差方差不相等之外,来自“准真值等价度量”的真值可能相差一个额外的常数(Novick & Lewis,1967),从而得到不相等的真值均值。

同属度量。在古典测量理论的框架下,这是限定最少的模型,它仅要求:旨在测量同一个现象的两个度量上的真值,完全相关即可(Jöreskog,1971b)。因此,在同属度量上,误差方差、真值均值、真值方差都可以不相等。

# 信度估计的方法

我们可以把信度刻画为一种误差的一种理论,或者更恰当地说,各种误差的各种理论。如前所述,测量误差具有各种来源,在一种参照系下被看成是误差的事物,在另一种参照系下则有可能不被看成是误差。毫不奇怪,误差的不同定义和概念会导致不同的信度估计方法。如前所述,"一个度量的唯一信度"这种说法是误导大家的原因。随着所涉及的误差来源不同,一定程度上,信度估值也会有所不同。因此,我们有必要在报告信度时,提供有关估计程序的充分信息,以便让大家能够了解所涉及的误差来源有哪些。下面,我们将在三大类别下,讲解最常用的信度估计方法。

## 测试—再测

从概念上和直觉上来说,"测试—再测"法是最简单的方式;如果我们把信度看成是测量的一致性或重复性,那么,这是最接近这种观点的方法。按照这种方法,我们采用相同的度量,把一群人测量两次,这样所取得的两组分数相关。所求得的相关系数,可看成是这个度量的信度估值。有些学者将这样求得的相关系数称为"稳定性系数"。

这种方法背后的假定是,两组观测值之间的相关关系,源自其背后不可观测的真值,这些真值是恒定的,而且,由于测量会出现随机误差,这两组分数之间也不会完全相关。很明显,对社会行为研究所感兴趣的大多数变量(例如,态度、动机、兴趣、成就)来说,这个假定并不现实。我们在前面已经论证过,就大多数(即使不是全部)这类属性而言,"再测"这个提法是一个错误命名。因此,我们仅需要指出一点即可:采用相同的度量依次测量被访者两次,这种做法也容易导致因"传导效应"而带来的偏差;所谓"传导效应"是指,第一次应答一组题器的动作本身就已经会影响第二轮给出的答案。一般来说,传导效应会倾向于高估从一个时期到另一个时期的稳定性,因而让信度估值产生通胀。

延长同一个度量的两次实施的间隔时间,这是尽可能减少传导效应的途径之一。尽管这能解决一部分问题,但它也会带来其

他问题。的确,两次实施的时间间隔越长,传导效应发生的概率就越小。但同样正确的是,时间间隔越长,在所研究的特质或特征上,个体发生真正变化的概率也越高。试举一例,一个较低的测试—再测相关系数,既可能说明一个具有较低信度的度量,也可能说明被测量的个体发生了真正的变化,或者两者兼而有之。我们的论点是,在测试—再测模型中,我们不可能区分一个度量的信度和稳定性。[①] 一般建议两次实施的间隔短一些(例如,一星期或两个星期),希望只触及随机误差而不触及真正的变化,这也是原因所在。因为测试—再测法具有严重的缺陷,我们的建议是,不予采用或者小心翼翼地加以采用。

## 等价形式

为了避免测试—再测法的一些内在问题,有人提出,我们可以采用旨在测量同一个现象的一个度量的两种不同形式,然后计算这两种形式所得分数的相关系数,作为信度的估值。在理想的情形下,这两个形式应当是平行的。但由于平行度量背后的各种限定性假定几乎无法满足(参见前文),我们经常也使用一些等价形式(或另类形式),它们或多或少有些偏离平行性。我们把这两个形式之间的相关系数看成是它们当中任一个的信度估值,并称为"等价形式信度"或"另类形式信度"。

等价系数不仅反映了信度,也反映了这两个形式测量相同属性的程度。而且,随着实施这两个形式的时间间隔不同,等价系数也可能反映了个体所经历的临时或持久变化。前者的一个例子是,个体的情绪在两次实施之间可能会发生变化;后者的一个例子是,个体在这期间获得了和被测量现象相干的信息。一般的建议是,等价形式间隔几天实施。这样的话,我们就有机会捕捉到两种误差:一种源自所用的特定形式,一种源自被访者所经历的临时变化。很明显,这样得到的系数可称之为"等价系数"或"稳定性系数"。

使用等价形式来估计信度,尽管它背后的理据具有直觉冲击

---

① 分离信度估值和度量的稳定性估值的各种程序已经得到开发(参见 Heise, 1969a; Werts et al., 1980; Werts et al., 1971; Wheaton et al., 1977; Wiley & Wiley, 1970),它们的最低要求是相同的度量实施三次。

力,但这个方法的用途有限,主要是因为在建构等价形式,以及在决定它们是否真的等价这两个方面所面临的困难。

## 内在一致性

在尝试两次接触相同的被访者时,更不用说,在确保他们能够给予合作,以便再次填答相同的度量(在测试—再测的情形下)或填答一个类似于前面填答过的度量(在使用另类形式的情形下)时,我们会遭遇各种实际困难。虽然这看起来并没有必要细述,但我们必须认识到,在这些情形下,被访者流失和自选择几乎是不可避免的,如果两次实施的间隔较长,情况更是如此。在本书第二部分的很多地方,我们将讨论被访者流失和自选择给一个研究的信度所带来的威胁。

一方面是由于上面提到的问题,一方面是由于对前面章节所提出的各种关注的回应,这些关注与测试—再测以及另类形式的方法有联系;有人提出了另一个信度概念,称做"内在一致性信度",它的基础是一个度量的一次性实施。

在讲解这个方法之前,我们有必要区分由单题器组成的度量和由多题器组成的度量,然后对这个区分作一个简短的点评。在本章的前面段落中,我们并没有提及这个区分,其原因在于,撇开效度问题不谈,测试—再测或另类形式的信度,它们都适用于这两种类型的度量。例如,我们既可以使用社会经济地位(SES)的单个指标(例如,收入),也可以在多个指标(例如,收入、教育、职业)的基础上形成一个复合 SES 指数。不管在哪一种情形下,只要我们能够获得两个时点上的应答,我们就可以估计出测试—再测信度;同理,我们也可获得另类形式的信度估值,只要它们的基础是单个题器或者是由多个题器组成的形式(更普遍的情形)。

为了具有实质性意义,一个复合分的基础必须是测量"相同现象"的各种题器。也就是说,对于一个属性、一个建构的一个度量的各种构成题器,我们期待它们的应答是内在一致的。这个概念正是"内在一致性信度"估值的基础。

### 裂半信度估值

采用内在一致性方法来估计信度的最早例子,就是后来众所周知的"裂半信度"估值。正如下面的讨论所表明的,这个方法具

有很大的局限性，因而我们应当尽量避免使用它；我们对它进行讲解，不但因为历史的缘故，而且因为它简单明了，在讲解其他内在一致性方法之前，我们可以把它作为一篇有用的导论。最重要的是，由于它应用广泛，我们相信，我们也应当关注它的各种局限。

我们可以把裂半法看成是另类形式的信度估值的一个变种。构成一个既定度量的所有题器，我们把它们分成两半，并把每一半看成是另一半的另类形式，这样，就没有必要去建构同一个度量的两种形式。就像另类形式的信度估值一样，这个度量的两半所得分数是相关的。但这个相关系数的基础是只有原来题器一半的一个度量。例如，如果原来的度量由 10 个题器组成，每一半由 5 个题器组成，那么，这两半之间的相关系数，就是一个长度为 5 个题器的度量的信度估值。为了估计出比每一半长一倍（即原度量的长度）的度量的信度，传统上，计算裂半相关系数的公式是"Spearman-Brown 公式"（1910 年，两人独立推导出该公式，因此得名）：

$$r_{xx}=\frac{2r_{1/2\ \ 1/2}}{1+r_{1/2\ \ 1/2}} \tag{5.14}$$

其中，$r_{xx}$为一个度量的信度，$r_{1/2\ \ 1/2}$ 是两半之间的相关系数。在上面的例子中，如果一个量表的两半（各自有 5 个题器）的相关系数是 0.62 的话，该量表的信度估值是：

$$\frac{2(0.62)}{1+0.62}=0.765$$

公式（5.14）是一个特例，Spearman-Brown 的通用公式：

$$r_{kk}=\frac{kr_{xx}}{1+(k-1)r_{xx}} \tag{5.15}$$

其中，$k$ 是该度量的长度扩大（或缩小）的倍数；$r_{kk}$是长度 $k$ 倍于该度量的一个度量的信度估值；$r_{xx}$是当前度量（即未改变长度之前的度量）的信度。请注意，Spearman-Brown 公式的基础是一个具有直观理据的预期：增加一个度量的长度会提高它的信度，减少它的长度会降低它的信度。

但是，Spearman-Brown 公式的效度建立在一个假定之上，即添加到一个度量或从它减去的部分，必须完全平行；即我们假定，它们具有相同的真值和相等的误差方差（参见前面对平行度量的讨论）。例如，如果我们将一个测试的长度翻倍，那么，我们就已经假

定,增加的那部分完全平行于原度量。如果这个假定不成立,Spearman-Brown 公式就会导致有偏估值。在本章的下面,我们将会说明,其他一些方法建立在更为宽松的假定之上,因而就很多(即使不是大多数)社会行为度量的信度而言,它们更接近现实一些。

在回到裂半法之前,我们将举例说明如何运用 Spearman-Brown 公式。假定我们建构了一个度量,它包含 5 个题器,它的信度估值是 0.40,那么,一个两倍于它的度量(即由 10 个题器构成的度量)的信度估值是多少？在本例中,$k$ 为 2,$r_{xx}$是 0.40。代入公式(5.15):

$${}_{kk} = \frac{2(0.40)}{1+(2-1)(0.40)} = 0.57$$

再举一例,现在假定一个度量包含 5 个题器,信度为 0.40,有一个研究者希望估计一个长度三倍于它(即 15 个题器)的度量的信度。此时,$k$ 为 3,$r_{xx}$是 0.40。代入公式(5.15)得到信度估值为 0.67。

我们需要注意两点:①一个度量的长度翻倍,它的信度并不会翻倍;②该度量的长度递增导致递减的信度收益。也就是说,随着我们不断增加该度量的长度,信度估值的增值将变得越来越小。其他条件保持不变的情形下,这个增值的具体幅度将取决于原度量的信度。虽然我们将在本章的后面讨论“期望达到的信度水平”这个话题,但在这里,我们应注意到,我们可以用公式(5.15)来求 $k$ 值,这样,为了达到一个信度水平,我们就可以估计出一个度量应当增加或减少多大长度。具体来说,

$$k = \frac{r_{kk}(1 - r_{xx})}{r_{xx}(1 - r_{kk})} \tag{5.16}$$

其中,$k$ 是该度量在长度上应扩大或减少的倍数,$r_{kk}$是期望达到的信度值,$r_{xx}$是当前度量的信度。

现在让我们回到估计信度的裂半法,请注意,便利是它的唯一优势。我们把现存的一个度量分成两半,就好像自己已经得到了这个度量的另类形式,而不必再去建构它的各种另类形式。因为这两半所得的分数是相关的,所以我们可以应用 Spearman-Brown 公式,估计出两倍于裂半的一个度量的信度。

众所周知,把一个度量分成两半的方法有很多。例如,对一个由 10 个题器组成的度量,存在 126 种不同的分法,可以把该度量分

成两半。[1] 很明显，各种裂半产生相等的裂半相关系数的概率十分低（在本章的后面，我们会举例说明，一个度量的不同裂半，如何会得出不同的相关系数）。这样的话，我们怎么才能判断一个裂半是"正确"的呢？答案是，这两个裂半应当平行，这与 Spearman-Brown 公式背后的假定一致。但说起来容易，做起来难，在把一个度量分成两半时，研究者一般采用就事论事的分法，最流行的两种分法是：①把所有奇数题器放到一半，所有偶数题器放在另一半（即奇偶法）；②把题器的前一半作为一部分，后一半作为另一部分（即前后法）。这两种分法的潜在问题是显而易见的。仅举一例，假定我们采用一种成就度量，其中，所有题器按照难度升序排列。在这种情形下，按照前后法，所得的两半明显不是平行的。

### $\alpha$ 系数

如前所述，估计信度的内在一致法的基础是一个概念，即量表的题器（或部分）测量相同的现象。一般而言，这意味着这些题器是同质性的。我们说"一般"，是因为就"这个概念的含义以及应如何测量同质性"而言，并没有达成一致意见。（Lord & Novick，1968：95。同时参见 Coombs，1950；Green et al，1977；Loevinger，1948；Scott，1960；Terwilliger & Lele，1979；Weiss & Davison，1981）在本章的后面，我们将再讨论这个问题。现在，我们仅指出，尽管对这个术语的含义还没有一致意见，但在讨论有关各种现象的不同度量时，"题器的同质性"仍然是一个具有直觉含义、十分诱人的术语。这些度量既可能推导自一个理论参照系，也可能因为各种实质基础（例如，各种特质、特征、属性）而引起我们的兴趣。

有关信度的内在一致估计法，人们提出过各种理论表述。在一定程度上，它们都涉及前面所提及的、广义的同质性；也就是说，它们均聚焦于一点：构成一个量表的各个部分或题器之间公有的东西是什么。解决这个问题的一般做法是，把度量中的每一个题器当做度量的一个部分，然后再考察它们之间的关系。请注意，我们可以把这种做法看做是裂半法的逻辑延伸，而且，我们还规避了把一个度量分成两半时所具有的任意性（参见前面有关裂半信度的讨论）。

---

[1] 可能的分法 $=(2n!)/2(n!)^2$，其中，$2n=$ 题器数，! 表示阶乘。

在研究每个题器之间的关系时,不同的理论取向起步于不同的假定,并采用了不同的分析方法,尽管如此,它们还是得到了基本上相同的信度估值。事实上,有些分析方法表面上看起来差异很大,但可以证明,它们所采用的公式在代数上是相等的。综上所述,我们将集中讨论一个公式,称为"$\alpha$ 系数",经常也称做"克伦巴赫阿尔法",这是因为克伦巴赫对这个方法的经典讲解和细化(Cronbach,1951),在估计致信度时,或许它是最常用的一个公式。在下一节,我们将点评其他方法,它们要么从属于 $\alpha$ 系数,要么和它相似。

$\alpha$ 系数具有很多种表达式,在代数上,它们都是相等的。最常用的一个表达式是:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right] \qquad (5.17)$$

其中,$k$ 为题器的数量;$\sum \sigma_i^2$ 是题器方差的总和;$\sigma_x^2$ 是总分(即复合分)的方差。

*复合分的方差*

在本章的前面部分,我们把一个观测值看做是由一个真值和一个误差项所组成的。由此可见[参见公式(5.5)及相关讨论],复合分的方差等于各成分的方差,加上各成分的协方差的两倍。对任何一个复合分,无论其成分有多少个,这都成立。也就是说,一个复合分的方差等于其各成分的方差和,加上所有可能的一对成分间的协方差之和的两倍:

$$\sigma_x^2 = \sum \sigma_i^2 + 2\sum \sigma_{ij} \qquad (5.18)$$

其中,$\sigma_{ij}$为题器 $i$ 和 $j$ 的协方差($i \neq j$)。

我们将在第 17 章中讨论协方差和相关系数。就当前的目的而言,我们只想表明,两个变量($X$ 和 $Y$)之间的协方差,可以表示为它们的标准差和它们之间的相关系数之积。$X$ 和 $Y$ 之间的相关系数可表示为:

$$r_{xy} = \frac{\sum xy}{N\sigma_x\sigma_y} \qquad (5.19)$$

其中，$\sum xy$ 是 $X$ 距离 $X$ 的均值的离差，和 $Y$ 距离 $Y$ 的均值的离差，两者的乘积之和；$N$ 是个案数，$\sigma_x$ 和 $\sigma_y$ 分别是 $X$ 和 $Y$ 的标准差。现在，两个变量 $X$ 和 $Y$ 之间的协方差可以表示为：

$$\sigma_{xy} = \frac{\sum xy}{N} \tag{5.20}$$

其中 $\sigma_{xy}$ 是 $X$ 和 $Y$ 之间的协方差，其他项已经在公式(5.19)中给出定义。代入公式(5.19)，则协方差可以表示为：

$$\sigma_{xy} = r_{xy}\sigma_x\sigma_y \tag{5.21}$$

关于公式(5.21)我们想作几点说明。当我们把 $X$ 和 $Y$ 的分数标准化之后(即 $z$ 分，参见第 17 章)，它们的标准差等于 1.00，公式(5.21)简化为 $r_{xy}$，即相关系数是标准分之间的协方差。由前文和公式(5.21)可见，当两个变量的相关系数为 0 时，它们之间的协方差也为 0。同理，若其他条件保持不变的话，两个变量之间相关系数越大，它们之间的协方差就越大。

有了上述考察之后，现在让我们回头讨论一下 $\alpha$ 的计算公式(5.17)。请注意，$k$ 是题器数。在公式(5.17)中，当题器数很大时，第一项[即$\frac{k}{k-1}$]接近于 1，因此可以忽略不计。反过来，我们将注意力集中到公式(5.17)中的关键项，即各题器的方差和与总方差之比。请注意，是 1 减去这个比值，因此，该比值越小，所得的信度估值越高。反之，该比值越大，信度估值就越小。如前所述，总方差等于各题器的方差和加上所有可能成对题器之间的协方差的两倍。利用这个概念，我们给出公式(5.17)的另一个表达式，它将有助于大家对信度的理解：

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum \sigma_i^2}{\sum \sigma_i^2 + 2(\sum \sigma_{ij})}\right] \tag{5.17'}$$

其中，$\sigma_i^2$ 是题器 $i$ 的方差，$\sigma_{ij}$ 是题器 $i$ 和 $j$ 的协方差。

由公式(5.17′)可见，构成一个总分的所有题器，只有当它们之间具有相关关系时，我们现在所讨论的这个比值，它在公式中所涉及的两项才会有所不同。在极端的情形下，当所有可能的成对题器之间的相关系数均为 0 时，总方差就会等于各题器的方差和。此时，各题器的方差和与总方差之比为 1.00，信度系数为 0。这个

结果也是合情合理的,因为题器之间缺乏相关性,就意味着它们之间没有共同之处,这和“内在一致性”的信度概念(即所有题器测量同一个现象)相矛盾。

上面讨论的是各题器之间的相关系数在估计 $\alpha$ 中的作用,现在我们将讨论一下协方差和方差各自在复合分(即总分)方差中的作用,并把它们进行比较,这样的话,我们就有可能更容易理解这个问题。如前所述,一组 $k$ 个题器两两配对,总共有 $\frac{k(k-1)}{2}$ 对。例如,当 $k$ 为 10 时,共有 45 个成对题器。这样,使用 10 个题器时,总分的方差基础就是 10 个方差(即各题器的方差和)和来自题器间的协方差(即每一对题器之间协方差之和的倍数)的 90 个元素。尽管这些协方差的具体幅度取决于题器之间的标准差和相关系数,但是,十分明显的是,和方差相比而言,随着题器数的增加,在决定总分的方差大小时,题器间的协方差所起的作用会越来越大。

上面这个示例的总分基础是 10 个题器,作为对照,我们现在来考察另一个示例,它的基础是一个由 40 个题器构成的量表。在此示例中,总方差等于 40 个题器的方差加上 780 对题器的协方差的两倍(即 1 560 个元素)之和。在 10 个题器的例子中,方差和协方差之比是 10∶90。在 40 题器的例子中,这个比值是 40∶1 560。归根结底,即使每对题器之间的相关系数都比较小,只要增加题器数,与各题器的方差和相比,量表的总方差也会大大增加。这样的话,$\alpha$ 值也会变得更大。一个结果是,在其他条件保持不变的前提下,题器之间的相关系数越大,它们之间的协方差就越大,由复合分所构成的一个量表,它的信度也就越大。我们还将在后面讨论到这些问题中的一些问题。

## 一个数据实例

现在,我们将用表 5. 1 中的数据实例,来讲解 $\alpha$ 的计算过程。正如我们用其他数据实例来进行练习时所做的那样,在这里,我们也是采用小型的数据集,以便我们采用手算或使用计算器,就可以完成所有的计算任务。我们认为,对读者来说,亲自动手来进行运算是十分重要的,因为在讲解示例时,我们只呈现了一部分带有综述性质的结果。例如,我们没有给出具体的计算过程,就报告了题器的方差(或者,它们之间的相关系数)。我们建议大家自己完成

所有的计算。如果您遇到任何困难,请参考第 17 章。我们在这里所遇到的各种统计量,在那里都有讲解。一旦大家掌握了基本的方法,我们就可以让计算机来完成运算任务。下面,我们将讨论和演示如何用计算机程序来估计信度。

表 5. 1 中所报告的数据实例,是 20 名被访者对一个量表的应答,该量表是由 4 个题器构成的。表 5. 1 还列出了总分,即每个被访者在 4 个题器上的得分之和。我们可以把这个量表看做是个人对堕胎或核裁军(仅举两例)的态度。我们要求被访者在 7 级量表上指明他们对每一个题器的赞同(或不赞同)程度,其中,1 表示"很不赞同",7 表示"十分赞同"。另一种表述方法是,我们可以把表 5. 1 中的题器看做是一个自填问卷的题器,我们要求被访者在 7 级量表上指明他们从事特定行为的频次(1 表示"从不",7 表示"经常")(在社会行为科学中,有关如何选择测量方法的讲解,包括对累加评分量表和自填量表的讨论,参见第 6 章)。如果大家能够从自己感兴趣的研究领域来挑选例子的话,这也会有助于大家理解。

表 5. 1 中倒数第二行中的数据是题器均值和总分均值。请注意,一个复合分的均值等于它的各部分均值之和。在本例中,总分均值等于 4 个题器均值之和(11. 10 = 2. 35 + 2. 65 + 3. 45 + 2. 65)。

**表 5. 1 4 个题器的得分和总分 ($N = 20$)**

| X1 | X2 | X3 | X4 | 总 分 |
|---|---|---|---|---|
| 3 | 2 | 6 | 3 | 14 |
| 3 | 5 | 6 | 2 | 16 |
| 1 | 2 | 1 | 3 | 7 |
| 5 | 2 | 3 | 2 | 12 |
| 1 | 2 | 2 | 5 | 10 |
| 6 | 5 | 7 | 5 | 23 |
| 5 | 3 | 5 | 6 | 19 |
| 1 | 1 | 3 | 1 | 6 |
| 1 | 1 | 3 | 1 | 6 |
| 5 | 6 | 6 | 3 | 20 |

续表

| X1 | X2 | X3 | X4 | 总　分 |
|---|---|---|---|---|
| 2 | 1 | 3 | 2 | 8 |
| 3 | 2 | 5 | 1 | 11 |
| 2 | 5 | 5 | 5 | 17 |
| 1 | 1 | 1 | 1 | 4 |
| 2 | 2 | 1 | 2 | 7 |
| 2 | 5 | 3 | 3 | 13 |
| 1 | 2 | 5 | 2 | 10 |
| 1 | 2 | 1 | 1 | 5 |
| 1 | 1 | 1 | 4 | 7 |
| 1 | 3 | 2 | 1 | 7 |
| $\sum$:47 | 53 | 69 | 53 | 222 |
| $\overline{X}$: 2. 35 | 2. 65 | 3. 45 | 2. 65 | 11. 10 |
| $\sigma^2$: 2. 628 | 2. 527 | 3. 847 | 2. 428 | 28. 690 |

这里,我们仅关注表 5.1 中的最后一行,它报告的是题器方差和总分方差。正如我们前面所讨论过的那样,只有当各个题器间的协方差为 0 时(即当题器之间没有相关关系时),总方差才等于各题器方差的总和。在本例中,题器方差的总和为 11. 43,而总方差是 28. 69。由公式(5. 18)可知,17. 26(28. 69—11. 43)等于题器间协方差之和的两倍。现在,我们用公式(5. 18)来计算总方差。我们很快就会发现,即使题器数很少,这也包含相当多的计算量。更不必说,我们通常不会采用公式(5. 18)来计算复合分的方差。现在我们之所以这样计算,是为了清楚地说明,构成一个复合分方差的各种成分究竟是什么。

在应用公式(5. 18)时,我们首先需要计算每个题器的方差,然后再计算每对题器之间的协方差。在本例中,我们需要计算 4 个方差和 6 个协方差,它们都呈现在表 5. 2 中(我们建议大家自己完成计算任务)①。

① 在本章,我们用 $N$(而不是 $N-1$)计算方差和协方差。

每个题器的方差列在表5.2的对角线上(请对照表5.1的最后一行)。每对题器间的协方差列在表格的对角线之上。例如,题器1与题器2之间的协方差是1.423。协方差是对称的指数,也就是说,题器1与题器2之间的协方差等同于题器2与题器1之间的协方差。因此,一般来说,我们也可以在对角线之下列出协方差。在表5.2中我们没有这样做,是为了腾出空间报告每对题器之间的相关系数,它们正是对角线之下的元素。下面,我们将用相关系数来说明前面讨论过的论点。

**表5.2 协方差(对角线上)、方差(对角线)和相关系数(对角线下)**
**4个题器,原始数据见表5.1**

| | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| X1 | 2.628 | 1.423 | 2.194 | 1.074 |
| X2 | 0.552 | 2.527 | 1.908 | 1.028 |
| X3 | 0.690 | 0.612 | 3.847 | 1.009 |
| X4 | 0.425 | 0.415 | 0.330 | 2.428 |

注:对角线的各项之和等于11.43,对角线之上各项之和等于8.636。

和表5.2不同的是,如果我们在一个表的对角线之上和之下均报告协方差时,那么,我们就会把它称为"协方差矩阵"(或"方差—协方差矩阵",请注意,我们可以把题器方差看做是它和自己的协方差)。采用一个协方差矩阵,我们就可以在构成这个矩阵的题器或成分的基础上,把所有元素加总,就可得到一个复合分的方差。把所有元素加总,就相当于把方差(即对角线的元素)和协方差之和的两倍(在一个协方差矩阵中,对角线下的三角是对角线上的三角的镜像)相加。不过,在表5.2的情形下,总方差(即28.70)等于对角线元素之和($2.628+2.527+3.847+2.428=11.43$)加上对角线之上元素之和的两倍[$2\times(1.423+2.194+1.074+1.908+1.028+1.009)=17.272$]。考虑到四舍五入所带来的误差,它与表5.1所报告的值相同。

在我们使用表5.2中所报告的结果来计算α值之前,让我们对该表对角线之下所列的4个题器之间的相关系数,作一个简短的讨论。请注意,所有相关系数都是正值,幅度中等,这表明这些题器具有共同的元素。具体来说,例如,题器3和题器4共享大约

0.11(即 $0.33^2$)的方差,题器 1 与题器 3 共享大约 0.48(即 $0.69^2$)的方差。因为所用度量的尺度单位会影响协方差的大小,所以一般我们很难阐释它们。如果采用每对题器的标准差,以及两者的相关系数,我们就可以计算题器间的协方差(参见公式 5.21)。例如,题器 2 与题器 3 的协方差为:

$$(0.612)\sqrt{2.527}\sqrt{3.847} = 1.908$$

同理,我们可以计算其他协方差。这个例子说明了我们前面的结论:在其他条件保持不变的前提下,题器间的相关系数越高,它们之间的协方差越大,因而,由它们所构成的复合分的方差就越大。

现在,让我们回到本节的主要任务——给表 5.1 中的 4 个题器计算 α 值。我们可以利用公式(5.17)或公式(5.17′)来完成这个任务。我们选择后者,因为它十分清楚地表明了构成量表的题器间的相关系数(以及协方差)所起的作用。将表 5.2 中的结果代入公式中,表 5.1 中的数据的 α 值等于:

$$\alpha = \frac{4}{4-1}\left(1 - \frac{11.43}{11.43 + 17.27}\right) = 0.80$$

因此,采用 α 值,在 4 个题器的基础上,这个度量的信度估值是 0.80。我们可以说,总分方差中的 0.80(或 80%)是稳定的(或系统性的)方差。请注意,如果所有题器间的相关系数均为 0 时,题器的方差和与总方差之比为 1.00(在括号中,分子和分母都是 11.43),这样,信度就是 0.00。在讨论 α 值背后的属性和假定之前,让我们先举例说明,如何把它应用到题器 01 编码的量表上。

## 01 编码题器的 α 值

对表 5.1 所列的示例数据而言,每个题器的赋值范围是 1 至 7。在测量态度、兴趣等时,我们经常使用这样的赋值方式。但有些类型的度量是由 01 编码的题器(也称做“二项题器”)构成的,即一个题器的赋值只能在两个值之中取一。测量成绩时的多选题是二项题器的最好例子。无论选项有多少,我们一般把多选题的答案赋值为正确(一般赋值为 1)或不正确(一般赋值为 0)。二项题器的其他例子包括只能在两个选项中挑选一个结果的题器(例如,同意/不同意,真/假)。我们将在第 6 章讨论有关这两种赋值的优

缺点问题。

首先,我们应注意,公式(5.17)是计算 $\alpha$ 值的一个通用公式,因此,它也应当适用于由二项题器所构成的度量。然后,我们将推导出公式(5.17)的一个特例,以便让我们有机会来:①讨论二项题器的一些特性;②说明实际上这是相当简单的计算,当我们用手工进行计算时,这一点十分有用;③更重要的是说明它和 Kuder-Richardson 20 相同,这是一个十分流行的公式(参见下文)。

尽管我们可以采用任意两个数来赋值一个二项题器(或变量)的答案,但最常用的还是01 编码。在考试或能力测验中,答对一道题赋值为1,答错一道题赋值为0,这也是十分自然的一种做法。在多选题上,如果我们把答对的题数相加,就会得到一个总分,这种做法就是一个例子。

二项题器的均值。当我们把二项题器的得分赋值为0和1时,该二项题器的均值就是得1分的人占总体的百分比。这是因为在计算均值时,我们将所有的1相加,然后除以总人数,由此得到得1分的人数百分比。因此,$p$ 通常用来表示二项题器的均值。顺便说一句,在进行成绩和能力测试时,$p$ 也被称为"题器难度指数",因为它表明了有能够通过一个题器的人占总数的百分比。请注意,$p$ 值越大,对被访者总体来说,则意味着题器越容易。

二项题器的方差。计算方差的通用公式(参见第17章)可以用来计算二项题器的方差。显然,我们可以看到,对二项题器来说,方差等于 $pq$,其中 $p$ 是得1分的人数百分比(即题器均值,参见上文);$q = 1 - p$,即得0分的人数百分比。二项题器的标准差为方差的平方根,即 $\sqrt{pq}$。请注意,当 $p = 0.5$ 时,二项题器的方差以及标准差达到最大值(分别是0.25和0.50)。

有了上述介绍,我们现在来考察二分计数测试的 $\alpha$ 值计算公式:

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum p_i q_i}{\sigma_x^2}\right] \qquad (5.22)$$

其中 $p$ 和 $q$ 在前面均有定义,而 $i$ 指题器 $i$。对照公式(5.22)和(5.17),我们注意到它们是相同的,除了在前面一个公式中,题器的方差和(括号中的符号 $\sum \sigma_x^2$)在这里由题器的 $pq$ 之和来表达,因为此处涉及的是二分计数的测量。简单来说,公式(5.22)是公

式(5.17)的一个特例。

例如,我们可以将表5.1中4个题器的分数重新进行二分编码。具体来说,我们把1,2,3的分数再编码为0,4至7的分数编码为1。需要强调的是,我们并不推荐这种做法;恰恰相反,在本书中,我们处处反对这种做法,因为它会导致信息流失,等等。例如,我们假定,表5.1中分数是对态度题的赞同程度。如果把得分如上所述编码为01的话,那么,我们就无法了解得分1至3的被访者之间的差异。同理,我们也无法区分得分4至7的被访者。实际上,我们是用这个例子来说明,对4个题器而言,和前面所得的信度估值相比较,信息流失会造成信度估值的降低。这个结果并没有出乎我们的意料,因为信度反映了被访者之间的系统性区分,而再编码把一些系统性区分抹掉了。总之,当题器或变量是以连续变量计分时,千万不要把它们再编码为01变量。只有当题器确实是二项题器时,公式(5.22)才适用。

表5.3记录了表5.1数据的再编码得分。请注意:①对每个题器而言,编码为1的个数等于得分的总和;②每个题器的均值等于在该题器上得分为1的被访者的比例(即$p$);③每个题器的方差等于$pq$;4)总分的均值等于各题器的均值之和(即$\sum p_i$,参见本章的前面有关部分)。

**表5.3 4个二项题器的得分和总分($N=20$)**

| X1 | X2 | X3 | X4 | 总分 |
|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 0 | 2 |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 4 |
| 1 | 0 | 1 | 1 | 3 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 1 | 0 | 3 |
| 0 | 0 | 0 | 0 | 0 |

续表

| X1 | X2 | X3 | X4 | 总 分 |
|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 1 |
| 0 | 1 | 1 | 1 | 3 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 |
| $\sum$ :4 | 5 | 8 | 5 | 22 |
| $\overline{X}$: 0. 20 | 0. 25 | 0. 40 | 0. 25 | 1. 10 |
| $\sigma^2$: 0. 160 | 0. 188 | 0. 240 | 0. 188 | 1 .490 |

和表 5.2 一样,我们在表 5.4 中报告了题器方差、题器间的相关系数和协方差,其中,题器方差的总和(即对角线的元素)是 0.776,对角线之上的元素之和的两倍为 0.716。因此,总分的方差为 1.492。

对照表 5.2 与表 5.4,可见,后面一张表中的数值明显小于前面的一张表。这是 01 编码的直接后果,它再次支持了我们反对这种做法的劝告。

**表 5.4　协方差(对角线上)、方差(对角线)和相关系数(对角线下)**
**4 个题器,原始数据见表 5.3**

| | X1 | X2 | X3 | X4 |
|---|---|---|---|---|
| X1 | 0. 160 | 0. 050 | 0. 070 | 0. 050 |
| X2 | 0. 289 | 0. 188 | 0. 100 | 0. 038 |
| X3 | 0. 357 | 0. 471 | 0. 240 | 0. 050 |
| X4 | 0. 289 | 0. 200 | 0. 236 | 0. 188 |

注:对角线的各项之和等于 0.776,对角线之上各项之和等于 0.358。

利用前面所得的数据,表 5.3 中,由 4 个题器所构成的一个度量的信度估值等于:

$$\alpha = \frac{4}{4-1}\left[1 - \frac{0.776}{1.492}\right] = 0.64$$

不出所料,如果将一个度量的题器进行 01 编码,和不编码之前相比,该度量的信度将下降:从 0.80 下降到 0.64。

## 内在一致性:理论取向和假定

诺维克(Melvin R. Novick)和刘易斯(Charles Lewis)曾指出,如果我们想让 α 值等于一个度量的信度,那么,构成这个度量的题器至少是准真值等价的。(Novick & Lewis, 1967)也就是说,我们假定,每个题器的真值之间只有一个常数的差异。若此假定不成立时,$\alpha$ 值就是信度的低端临界估值。换言之,当题器至少不是准真值等价时,$\alpha$ 系数会低估一个度量的信度。

在前一节,我们介绍了一个 $\alpha$ 系数的形式,它对二项题器的计算特别有用,详见公式(5.22)。这个公式还有另一个名称"Kuder-Richardson 20",简称"KR-20",它是以首创该公式的两个作者的姓氏命名的[20 不过是库德(G. F. Kuder)和理查德森(M. W. Richardson)在一篇论文中所使用的公式编号,参见 Kuder & Richardson,1937]。请注意,尽管 $\alpha$ 系数与 KR-20 相同,但后者的假定比前者更严格。具体来说,KR-20 的假定是,构成一个度量的所有题器都是平行的。如前所述,这意味着所有题器的真值都是相同的,而且误差也相同。毋庸赘言,相对于 $\alpha$ 系数所作的假定而言,这个假定离现实更远。但需要唠叨一下的是,采用 $\alpha$ 系数和 KR-20,对信度的估计是相同的,因为它们两个的公式相同。

当题器方差是相等时,α 可以用下式来表达:

$$\alpha = \frac{k\bar{r}_{ij}}{1 + (k-1)\bar{r}_{ij}} \qquad (5.23)$$

其中 $\bar{r}_{ij}$ 等于 $k$ 题器间平均的相关系数。

众所周知,公式(5.23)被称为"信度的平均 $r$ 估值"。请回想一下,标准分(z 分)的均值为 0、标准差为 1,由此可推知,当我们把

题器的分数标准化后,公式(5.23)的结果就是 $\alpha$ 值,有时,它被称为"标准化题器的 $\alpha$ 值"(参见下面有关计算机的信度估计)。

我们引入公式(5.23),并不是因为我们相信,"题器方差相等"是一个比较现实的假定,更不是因为我们建议把题器分数标准化,而是为了把估计信度的 $\alpha$ 系数法与 Spearman-Brown 法联系起来。在本章的前面段落中,我们已经指出,估计信度的 Spearman-Brown 法具有更严格的假定,即构成一个度量的题器或成分是平行的,参见公式(5.15)。可以证明,公式(5.23)是更广义的 Spearman-Brown 公式(5.15)的一个表达式。但因为我们已经证明公式(5.23)是 $\alpha$ 系数的一个特例,由此可推,它的假定比公式(5.15)更严格。我们再次碰到两个相同的公式,但假定却不相同的情形。

如前所述,我们已经证明,估计信度的裂半法是 Spearman-Brown 法的一个特例。同时,我们也注意到,裂半法具有严重的缺陷,因为一个度量可以用不同的方式分成两半,从而有可能造成众多不同的信度估值。因为 $\alpha$ 系数的概念基础是,把一个度量分裂成和它的题器数量一样多的部分,所以,它避免了这个问题。而且,克伦巴赫(Cronbach,1951)证明,对一个既定度量而言,$\alpha$ 系数是它的所有裂半信度系数的均值。

特赖恩(Tryon, 1957)曾批判过古典测试理论及其变种,认为它们是建立在高度严格、不切实际的假定之上,并提出另一种理论取向——域抽样(除了特赖恩的论文之外,下列文献对域抽样也有很好的探讨,参见 Ghiselli et al. , 1981, 第8章; Nunnally. 1978, 第6章)。我们也可以从它推导出一个和 $\alpha$ 系数相同的公式(5.17),但和古典测试理论及其各种变种所推演的各种公式相比,这个公式对社会科学中所采用的测量类型,具有更切实际的概念,因而也具有更宽松的假定。

当不同的理论取向均可以推导出相同的公式时,我们肯定会产生疑问,对估计信度而言,它们具有什么(如果有的话)实践含义?只有当我们依次考察度量的长度和同质性与内在一致的信度之间的关系时,我们才能更好地回答这个问题。

## 测试长度和内在一致的信度

当我们把一个既定测量工具的题器分数看做是由真值和随机误差两个部分构成时,那么,有理由推断,当增加测量相同现象的

题器时,这个工具的信度就会提高。例如,Spearman-Brown 法背后的理据就是如此。

但我们应当认识到,内在一致的信度估值是构成一个测量工具的题器数以及题器之间的相关关系的互动结果,参见公式(5.17)、(5.17′)和(5.18),以及和公式相关的详细讨论。一个度量的不同题器之间的相关系数比较低,这在一定程度上表明,这些题器看起来是在测量不同的事物。即使在这种情形下,当题器数增加时,相对于题器的方差之和而言,总方差增加得更快。换言之,当题器数足够大时,可以证明,即使当构成一个度量的各种题器之间几乎没有共性,这个度量仍可以具有较高的内在一致的信度。我们还可以证明,若题器间的相关系数保持不变,当题器数趋向于无穷大时,$\alpha$ 趋向于 1。

然后,也有可能出现另一种情形,题器数虽然较少,但只要题器间的相关系数较高,我们仍然能够得到较高的信度。例如,由 3 个题器构成的一个度量,如果平均的相关系数等于 0.50 的话,它的 $\alpha$ 系数仍有可能会达到 0.75。当题器间的平均相关系数等于 0.25 时,若想得到同样的 $\alpha$ 系数,题器数就必须达到 9 个;若平均相关系数等于 0.10 时,题器数就需要 27 个。

## 同质性及内在一致的信度

前面我们已经指出,大家对“同质性”意义并没有形成一个共识。罗兹布姆(Rozeboom, 1966: 321)曾对这个问题有过深入探讨,并提出了估计同质性的公式,他主张,同质性“本质上就是一种相关系数的均值”。同时,洛德(Fredrick M. Lord)和诺维克也认为:“可见,一个同质的测试就是它的所有成分都在真值的意义上‘测量同一个事物’。”(Lord & Novick, 1968: 95)不过,洛德和诺维克并没有给出一个标准,界定什么是“测量同一个事物”。很多研究者(参阅 Green et al, 1977)认为,这就是指“单维性”或“单因子”。这些术语都来自因子分析的领域——这是第 22 章和第 23 章将要讨论的主题(有关因子分析的直观介绍,参见第 4 章)。

当我们把同质性理解为单维性时,这意味着,如果我们想把一个测量工具看做是由各种同质的题器所组成,那么我们就必须证明,一个公因子就足以解释题器间的相关关系。由此可推,一个测量工具也是内在一致的。但反过来不一定成立。一个内在一致的

测量工具并非必然是同质的。

克伦巴赫(Cronbach, 1951: 320)并没有把 $\alpha$ 系数局限在一种单维的测量工具上:"$\alpha$ 系数估计了题器间的所有公因子所带来的方差占一个测试的总方差的比例。"也就是说,它报告的是,测试分数在多大程度上取决于公因子和组因子,而不是取决于随题器不同而异的因子。

我们需要注意两点:①克伦巴赫使用"公因子"这个术语,也用来指第一个因子,而且,就目前的关注点而言,最重要的是②即使题器间的关系背后不存在一个公因子,$\alpha$ 系数仍可以很高。当题器间的关系背后存在两个或多个公因子或组因子时,就会出现这种情形。在前面,我们讨论过题器间的平均相关系数和题器数之间的互动对决定 $\alpha$ 系数的影响,至少在一定程度上,它们应当让我们明白出现上述情形的缘由。最明显不过的地方是公式(5.23),由此可见,$\alpha$ 系数的基础是题器间的平均相关系数。当一个测量工具是由不同组的题器组成时,若题器数既定,为了达到一个较高的 $\alpha$ 系数,题器间的平均相关系数可能仍需要较大的水平(这取决于一组题器内部的各题器间的关系强度以及题器组数)。进一步说,即便题器间的平均相关系数很小,如果题器数足够大的话,$\alpha$ 系数也会很高(参见前面对这个论点的讨论和举例说明)。

上述讨论给我们带来的最重要的经验是,我们不应当把 $\alpha$ 系数看做是一个测量工具的同质性指数。即使是粗粗浏览一下测量方面的文献和实质性的研究报告,十分明显的是,很多人都把内在一致的信度估值作为一个测量工具的同质性或单维性指数(有关示例和一篇精彩的讨论,参见 Green et al,1977)。

### 同质性、内在一致性和效度

我们可以采用构成一个度量的各个题器测量同一个事物的程度,来预测一个总分的效度。不过,正如前面所讲解的那样,即使不存在一个公因子,或者是一个因子不能解释一个测试的总方差中的大部分时,我们仍可以得到很高的 $\alpha$ 系数。因此,具有高信度的总分却有可能具有可疑的效度。

有意思的是,我们还会碰到一些情形,其中,研究者会有意设计由异质的题器所构成的度量。而且,他们不仅会指出,他们对自己的度量具有低内在一致信度估计的期望,而且把低信度估值(例

如,低 $\alpha$ 系数)作为自己成功建构这个度量的证据。对效度以及它和信度的关系来说,这个话题极其重要,因此,让我们考察一个例子。

在为自己的内外控制的度量选择题器时,米舍尔(Walter Mischel)等人指出,因为他们有意"试图在尽可能多样的情境和结果中进行抽样……不可避免的是,我们预期这个度量的内在一致信度会比较低(Mischel et al, 1974: 267)"。简言之,米舍尔等人建构了一个14题器的量表,得到了3个分数:一个总分,一个正子量表分,一个负子量表分。把裂半相关系数代入 Spearman-Brown 公式后,得到下列信度估值:正子量表为0.14,负子量表为0.20,而总量表为0.04。[①]

有些题器是服务于单纯的预测,有些则是服务于测量一个建构,这两者应当区别开来。在没有理论的语境下,对预测目的而言,把异质的题器得分加总,或许是有用的,但或许分别使用它们(例如,在回归分析中)会更有机会提高预测力。即便如此,像米舍尔等人那样,把异质的题器相加,并主张这个总分反映了一个建构,从术语上来看,这就是一个自相矛盾。[②]

当然我们还可以把"控制点"(一个米舍尔等人想要测量的建构)看做是一个多维概念。相应地,我们就有必要建构多个子量表来测量这个建构。不过,为了让子量表的分数有意义,构成子量表的题器必须是同质的(我们已经在第4章讨论过这个话题)。

最后,当我们的关注点是在测量一个建构时,相对较低的内在一致的信度,起着反面证据的作用。也就是说,正如我们反复强调的那样,虽然内在一致的信度较高,这并不是一个度量具有同质性的证据,但是,内在一致的信度较低,却可以是这个度量不具有同质性的证据。

## 选择信度估计时需要考虑的因素

考虑到有不同的方法可以估计信度,同时考虑到信度的不同界定却推导出相同的公式这个事实(参见前面有关 $\alpha$ 系数及其变

---

① 就此目的而言,裂半和其他内在一致的信度估值都是无效的。

② 参阅本章后面有关不可靠度对相关系数的逆效应,在那里,我们会进一步评述这个研究。

种的讨论),我们有必要考察一下选择一个特定方法的标准问题。它具有两个相关联的问题:①待测量的属性的性质;②就被测量的属性而言,十分重要而需要加以考量的误差来源。

如前所述,估计信度的不同方法,它们的主要差异在于,考察哪些误差来源,或者把变异度中的哪些成分看做是随机误差。例如,在测试—再测法中,从一个测试周期到另一个测试周期之间的所有变化,我们都把它们看做是随机误差。对相对稳定的属性(例如,智力)而言,这是站得住脚的,但对相对不稳定的属性(例如,情绪)而言,却是站不住脚的。

另一方面,信度的内在一致估值主要是针对因内容的抽样而带来的误差,即一个度量的题器对被测建构的概念域的代表程度。可见,从测量的域抽样概念出发,显而易见,这种方法是最有意义的。除了内容抽样所带来的误差之外,内在一致的信度还涉及单一情境下临时波动(例如,疲劳、情绪、注意力和猜题)所带来的误差。

毋庸赘言,我们并不是说上述评述穷尽了各种因素,而只是指出,在选择估计信度的不同方法时,我们所要考察的各种因素中的某几个因素(更详细的讨论参见 Nunnally,1978:第 7 章)。在测量建构时,通常,$\alpha$ 系数是被选中的估值,在特定情境、特定建构下,我们也可能会认为其他估值更切合一些测试结构的信度估算方法,而其他信度估算方法则与给定情境下的给定结构更为相关。大多数情形下,在决定采用何种信度估值时,研究过程中的理论和实践考量才是关键。

## 计算信度的计算机程序

第 16 章将对统计分析的计算机软件包进行一个总体介绍,同时,也将说明我们选择特定程序包的标准。十分有趣的是,只要是我们能够想象得到的统计分析,几乎总会有相对应的综合性软件包,尽管如此,有些软件包(例如,BMDP 和 MINITAB)却还是没有一个估计信度的程序。直到最近,SAS 才在自己的 PC 版本中,把 $\alpha$ 系数的计算作为 PROC CORR 命令中的一个选项,它的主机版本仍没有包括这个选项。这种遗漏看起来令人迷惑不解,特别是当我们认识到,和那些包含在主机版本中的极端复杂程序相比,信度程

序是很简单的编程任务。难道这种遗漏植根于这个软件包的消费者们对信度估值缺乏兴趣(知觉到的或表达出来的)?

在下面的演示中,我们将采用 SPSS 的 RELIABILITY 命令。

我们假定您对 SPSS 的控制语言有所了解,因此,我们仅对输入语句作简要的评注。如果您不熟悉这个软件包,或者不熟悉一般的计算机程序,您也可以阅读第 16 章的相关章节。

## 数 据

本次分析所使用的示例数据来自表 5.5,它的布局和表 5.1 相同;不同之处在于,表 5.5 是 20 个被访者在 10 个题器上的应答数据,表 5.1 则是 20 个被访者在 4 个题器上的应答数据。而且,正如表 5.5 的注释所指明的那样,该表中的前 4 个题器和表 5.1 中的前 4 个题器相同。随着分析的进行,我们选择这样做的理由就会逐渐清晰起来。为了方便和计算机输出的比较,在表 5.5 中,我们还报告了均值和方差。

**表 5.5 10 个题器的得分和总分 ($N=20$)**

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | 总 分 |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 2 | 6 | 3 | 3 | 3 | 3 | 5 | 2 | 5 | 35 |
| 3 | 5 | 6 | 2 | 1 | 2 | 5 | 6 | 2 | 3 | 35 |
| 1 | 2 | 1 | 3 | 2 | 3 | 1 | 2 | 2 | 1 | 18 |
| 5 | 2 | 3 | 2 | 1 | 5 | 1 | 2 | 2 | 2 | 25 |
| 1 | 2 | 2 | 5 | 1 | 1 | 1 | 3 | 3 | 1 | 20 |
| 6 | 5 | 7 | 5 | 3 | 3 | 2 | 2 | 1 | 5 | 39 |
| 5 | 3 | 5 | 6 | 5 | 5 | 7 | 5 | 6 | 6 | 53 |
| 1 | 1 | 3 | 1 | 4 | 5 | 7 | 2 | 2 | 2 | 28 |
| 1 | 1 | 3 | 1 | 2 | 1 | 2 | 1 | 3 | 5 | 20 |
| 5 | 6 | 6 | 3 | 5 | 5 | 3 | 5 | 3 | 1 | 42 |
| 2 | 1 | 3 | 2 | 1 | 2 | 1 | 1 | 2 | 3 | 18 |
| 3 | 2 | 5 | 1 | 1 | 3 | 3 | 1 | 3 | 3 | 25 |
| 2 | 5 | 5 | 5 | 3 | 1 | 3 | 1 | 5 | 1 | 31 |

续表

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 | 总 分 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 2 | 1 | 3 | 2 | 1 | 14 |
| 2 | 2 | 1 | 2 | 2 | 3 | 2 | 2 | 1 | 3 | 20 |
| 2 | 5 | 3 | 3 | 2 | 3 | 3 | 5 | 3 | 7 | 36 |
| 1 | 2 | 5 | 2 | 3 | 1 | 2 | 3 | 2 | 3 | 24 |
| 1 | 2 | 1 | 1 | 1 | 1 | 2 | 5 | 1 | 2 | 17 |
| 1 | 1 | 1 | 4 | 1 | 5 | 1 | 3 | 5 | 2 | 24 |
| 1 | 3 | 2 | 1 | 7 | 5 | 3 | 1 | 2 | 5 | 30 |
| $\sum$ :47 | 53 | 69 | 53 | 49 | 59 | 53 | 58 | 52 | 61 | 554 |
| $\overline{X}$:2.35 | 2.65 | 3.45 | 2.65 | 2.45 | 2.95 | 2.65 | 2.90 | 2.60 | 3.05 | 27.70 |
| $\sigma^2$:2.628 | 2.527 | 3.847 | 2.428 | 2.747 | 2.348 | 3.127 | 2.690 | 1.740 | 3.247 | 93.710 |

注：前 4 个题器和表 5.1 相同。

**输入**

```
SET LISTING = 'T55SPS. LIS'.
TITLE RELIABILITY ESTIMATES. DATA OF TABLE 5.5.
DATA LIST FREE/X1 TO X10.
BEGIN DATA.
3 2 6 3 3 3 3 5 2 5[第一个被访者的数据]
........................
1 3 2 1 7 5 3 1 2 5[最后一个被访者的数据]
END DATA.
LIST.
RELIABILITY VARIABLES = ALL /
  SCALE(ATTIT) = ALL /
  SCALE(ATTIT) = ALL /MODEL = SPLIT /
  STATISTICS ALL /SUMMARY ALL.
```

**注解**

这个命令运行在 PC 版的 SPSS 上。参阅第 16 章，我们将解释 SPSS 的主机版本和 PC 版本的异同之处。

我们使用的是小数据集，遵循惯例，我们把数据直接作为命令

文件一部分。由此可见,输入格式是 Free,这种输入格式仅需要用至少一个空格或逗号把数据值分隔。X1 TO X10 这个设定是为 10 个变量命名(在这里就是 10 个题器)。正如第 16 章所解释的那样,在这个示例中,我们加入了 LIST 命令,以便让原始数据成为输出的一部分。

RELIABILITY 程序要求,分析中所用到的变量,必须在 VARIABLES 子命令后加以指明,它也接纳关键词 ALL。之后,我们就有可能使用所有题器及其组合来进行分析。构成一个量表或子量表的每组题器,都可以在 SCALE 子命令下加以指明。在本例中,所有题器都用于一个量表中,它的名字 ATTIT(在括号中)表示态度。

RELIABILITY 提供了估计信度的几种模型。当 MODEL 子命令并没有指定一个模型时,如本例中的第一个 SCALE 子命令后,默认模型是 ALPHA。出于比较和演示的目的,我们也用 MODEL = SPLIT(即裂半信度估值)来分析数据。请注意,SPSS 采用的前后裂半法,它将前$\frac{k}{2}$个题器作为前半部分,剩余题器作为后半部分。

我们可以 STATISTICS 和 SUMMARY 子命令下设定各种题器和量表的统计量(例如,均值、标准差、题器之间的方差—协方差矩阵、题器之间的相关系数矩阵、题器—总分的相关系数)。ALL 也是一个可接纳的关键词。

**输出**

RELIABILITY COEFFICIENTS 10 ITEMS

ALPHA = 0.787 1 STANDARDIZED ITEM ALPHA = 0.784 6

**注解**

$\alpha$ 系数等于 0.79 表明这个量表的方差当中,79% 是系统性的。我们可以代入公式(5.17)求得这个值。采用题器的标准分,我们可以估计标准化的题器 $\alpha$ 系数;或者,以等价的方式,采用平均题器间相关系数,代入公式(5.23)求得。作为一个练习,我们建议您使用表 5.6 的对角线下的数据,来计算标准化题器的 $\alpha$ 系数。本例中,题器方差彼此相近,$\alpha$ 系数的两种估值几乎相等(参见与公式(5.23)有关的讨论)。

**输出**

CORRELATION BETWEEN FORMS = 0.543 7

EQUAL LENGTH SPEARMAN-BROWN =0.704 4

**注解**

上面的输出节选自 MODEL = SPLIT 命令(参见输入命令)所得的输出结果。量表的两半之间的相关系数为 0.543 7。代入 Spearman-Brown 的裂半公式(参见公式(5.14)以及相关讨论),求得这个量表的信度估值为 0.704 4。如前所述,当我们设定 MODEL = SPLIT 时,SPSS 会把一个测试分成前-后两半。作为练习,我们建议您进行奇偶裂半,并计算裂半相关系数,计算的结果是 0.723。把它代入 Spearman-Brown 公式(5.14),求得信度估值是 0.839。请注意,这两个裂半信度估值之间存在相当大的差异,这进一步强化了我们的建议:不要采用裂半法来估计信度。

## 二项题器的 α 系数

我们在前面提到,当题器为 01 编码时,α 系数也适用[参见公式(5.22)以及相关讨论],而且,在这种情形下,α 系数公式和 KR-20 得到相同的结果。作为练习,您可以将表 5.5 中的数据进行 01 编码。这项转换工作没有必要用手工来做,RECODE命令就可以完成,它出现在 DATA LIST 命令之后。

RECODE X1 TO X10 (1 THRU 3 =0) (4 THRU 7 =1)

前面所列的所有输入语句,都不会受到影响。如果您用这种形式计算表 5.5 的数据,求得的 α 系数是 0.718。我们反复强调,不要把连续题器分进行 01 编码。这里仅仅是演示,为了说明,当题器分被 01 编码时(在成绩或能力度量中,常常这样做),SPSS 的 RELIABILITY 程序求得的结果,和 KR-20 相同。

## 对表 5.5 数据的详细考察

如前所述,表 5.5 中的前 4 个题器与表 5.1 中的题器相同。对这 4 个题器而言,前面的 α 系数估计为 0.86。增加 6 个题器之后,较长量表的信度却稍微降低了一点(0.79,参见前面)。因此,增加的 6 个题器与前面的 4 个题器,不是准真值等价的;它们没有测量相同的事物(参见本章前面的详细讨论)。

如果被访者的数量足够大的话,做一个因子分析是有用的,它可以让我们了解到对题器应答背后的结构。但对当前的目的而言,一个不太严谨的方法就足够了。实际上,我们将要做的工作

是,对 10 个题器的相关系数矩阵进行“目测”因子分析。

表 5.6 对角线下是 10 个题器的相关系数矩阵,它是从 SPSS 的输出中复制的。为了便于观察,我们把表中的变量排成两组。第一组由题器 X1 至 X4 组成,第二组由题器 X5 至 X7 组成。请注意,第一组的题器间相关系数从 0.33 到 0.69,第二组的题器间相关系数是从 0.36 到 0.51。跨组的题器间相关系数一般低于同一组内的题器间相关系数,有一些例外(例如,X1 与 X6 之间的相关系数)。同时,X8 到 X10 和前面提到的两组题器间的相关系数一般也比较低,也有少数例外(例如,X4 与 X9 之间的相关系数)。简言之,两组题器看起来是在测量不同的事物。

需要提醒大家的是,我们的目的是对因子分析进行一个最粗略的逼近。我们所做的工作也表明,即使是在较小的相关系数矩阵中,尝试发现题器或变量之间的一种模式,也是相当困难的。因此,它也表明了诸如因子分析方法的优点,在第 22、23 章中,我们将讨论它们。

从前面的分析中,我们了解到,题器 X1 至 X4 的 $\alpha$ 系数等于 0.80,题器 X5 至 X7 的 $\alpha$ 系数等于 0.72。利用表 5.6 中对角线和对角线之上的数据,代入公式(5.17)中,我们就相对容易地计算出这两个 $\alpha$ 系数。如果大家希望通过 SPSS 的 RELIABILITY 来计算这两个 $\alpha$ 系数,所要做的事情,就是在前面的输入语句中,增加下列两个子命令。

```
SCALE(ATTIT1) = X1 TO X4/
SCALE(ATTIT2) = X5 TO X7/
```

为了便于识别,我们把第一组命名为 ATTIT1,把第二组命名为 ATTIT2。在实际研究中,我们可以采用其他标签,它们反映出每组题器所期望测量的事物。在实际研究中,我们还需要进行其他决策。例如,我们需要决定是否舍弃题器 X8 至 X10,是否给第二组增加题器,等等。我们此时的目的仅仅是想说明,完全依赖 $\alpha$ 系数,让我们并不拥有充分的信息,甚至会误导我们。很多研究者认为,10 个题器的 $\alpha$ 系数等于 0.79,这是同质性或单维度的指标。但我们非常粗略的考察已经表明,这是一个错误的结论。

表 5.6 协方差(对角线上)、方差(对角线)、相关系数(对角线下)
10 题器,原始数据在表 5.5

| | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 2.628 | 1.423 | 2.194 | 1.074 | 0.543 | 1.018 | 0.673 | 0.635 | 0.190 | 0.733 |
| X2 | 0.552 | 2.527 | 1.908 | 1.028 | 0.908 | 0.182 | 0.728 | 1.015 | 0.260 | 0.517 |
| X3 | 0.690 | 0.612 | 3.847 | 1.009 | 0.948 | 0.072 | 1.457 | 0.795 | 0.330 | 1.077 |
| X4 | 0.425 | 0.415 | 0.330 | 2.428 | 0.357 | 0.182 | 0.277 | 0.515 | 1.210 | 0.268 |
| X5 | 0.202 | 0.344 | 0.291 | 0.138 | 2.747 | 1.273 | 1.507 | -0.005 | 0.380 | 0.977 |
| X6 | 0.410 | 0.075 | 0.024 | 0.076 | 0.501 | 2.348 | 0.982 | 0.145 | 0.430 | 0.403 |
| X7 | 0.235 | 0.259 | 0.420 | 0.101 | 0.514 | 0.363 | 3.127 | 0.965 | 0.710 | 1.068 |
| X8 | 0.239 | 0.389 | 0.247 | 0.202 | -0.002 | 0.058 | 0.333 | 2.690 | 0.160 | 0.505 |
| X9 | 0.089 | 0.124 | 0.128 | 0.589 | 0.174 | 0.213 | 0.304 | 0.074 | 1.740 | 0.170 |
| X10 | 0.251 | 0.181 | 0.305 | 0.095 | 0.327 | 0.146 | 0.335 | 0.171 | 0.072 | 3.247 |

注:本表是 SPSS 的 RELIABILITY 程序输出的两个部分(COVARIANCE MATRIX 和 CORRELATION MATRIX)的拼接。对角线元素之和是 27.239,对角线之上元素之和是 93.710。对两组题器的讨论参见正文。

# 几个话题

我们将选择与信度关联的几个话题进行讲解,并结束本章的讨论。对它们的涉猎深浅不同:有的讲解比较细致,有的只是一笔带过,目的只是让您熟悉它们,然后指出详细讨论的文献出处。请记住,我们使用广义的"信度"术语(例如,在各种公式中),"是哪一种信度估值?"对这个问题的回答取决于情境的特殊性。

## 信度的"标准"

一个度量的信度应当有多大? 显而易见,其他条件保持不变时,信度越大越好。由此推知,对一个幅度既定的信度系数,在一些情形下我们认为它是可接纳的;在另一些情形下,我们则可以认为它是不可接纳的。在判断一个信度系数是否可接纳的各种考量中,最重要的一个考量是和分数基础上的决策类型以及决策的可能后果有关联。有鉴于此,很多学者提出了关于可接纳的信度系

数的标准或最低水平的经验规则。(例如,Nunnally,1967: 226, 1978: 245-246;Thorndike & Hagen,1977: 92-94)例如,有一种观点认为,在研究的早期阶段,我们应当容忍相对较低的信度系数;当这个度量开始用于决定不同群体之间的差异时,我们就需要较高的信度,如果分数是用于对个体进行重要决策(例如,选择和任用决策)时,高信度就是基本的要求。

尽管这些"把特定值作为信度的标准"的建议都建立在坚实的推理之上,但我们仍然质疑这是不是一种明智之举。因为这些标准会逐渐获得自己的生命,它们经常得到应用,形成它们的观念源头却没有得到重视。一个典型的情形是,始作俑者显然已经改变了自己的观点,但他们所提出的标准却继续得到应用。一个切题的案例是遵循纽纳利所提出的经验规则,他(Nunnally, 1967: 226)认为:"在对预测测试或一个建构的假设度量进行研究的早期阶段,我们可以采用只有中等信度的测量工具,以节省时间和精力;就此目的而言,0.60 或 0.50 的信度就足够了。"纽纳利(Nunnally, 1978: 245)重复了上述观点,但有一点修正:"0.7 或以上的信度就足够了。"在使用信度系数相对较低的度量时,研究者似乎有一种需求,即引用一个权威来"佐证"自己的做法;这时,他们就可能倾向于引用那些最适合自己目的的"标准"。例如,卡普兰(Caplan et al, 1984: 306)等人写道:"就本研究目的而言,$\alpha$ 值等于或大于 0.50,我们判断它就是恰当的",此时,他们引用的是纽纳利(Nunnally, 1967)。更有趣的是,埃利斯(Ellis,1988)引用纽纳利(Nunnally, 1978)的研究,并得出结论:他的所有度量都达到了"可接纳"的信度水平(Ellis, 1988: 685),尽管事实是它们当中的几个低于引文中纽纳利所设定的标准 0.7。

有时,同事或学生会来问我们,在使用自己的信度估值时,他们应当引用什么文献来佐证;我们开玩笑地说,如果他们的信度估值在 0.7 左右时,那么就引用纽纳利(Nunnally, 1978),但如果信度估值只有 0.5 左右时,那么就引用纽纳利(Nunnally, 1967)。在更正式的场合,我们就会向提问者指出,这不是一个权威宣布一个特定信度系数是不是清真食品的问题。为了加强效果,我们还讲了一个故事,一个妇女请教一个拉比①,鸡肉是不是清真食品。拉比

---

① 犹太教祭司。

看了看,然后开始闻。这个举动让这个妇女感到吃惊,她觉得在这种情形下,拉比的举动有点特别。她禁不住再问道:"拉比,这是清真食品吗?"拉比回答道:"是的,只不过臭了。"使用一个信度系数为0.5或其他什么值的度量,是清真食品吗?当然是!一个0.5的信度系数臭了吗?没有任何一个权威来源可以回答这个问题。相反,在特定的研究情形(例如,分数的用途、研究的成本)下,我们需要决定自己能够容忍的误差量究竟有多大。

## 预测真值

在本章的前面我们曾指出,我们把观测值看做是由真值和误差两个成分组成的(参见公式(5.1)及其相关讨论)。应用线性回归分析(参见第17章)的概念和假定,以观测值预测真值的方程是:

$$\begin{aligned} T' &= a + bX \\ &= (1 - r_{xx})\bar{X} + r_{xx}X \end{aligned} \tag{5.24}$$

其中 $T'$ 为预测的真值,$a$ 为截距,$b$ 为回归系数,$X$ 为观测值,$X_{rr}$ 为信度系数,$\bar{X}$ 为一组观测值的均值。

根据您的背景知识,或许您想和第17章的相关章节一起来学习本段落。就当前的目的而言,请大家注意真值对观测值的回归方程的这个特殊表达式,即公式(5.24)的第二个表达式。由此可见,第一项是截距($a$),信度系数为回归系数($b$)。

在预测真值时,这两项如何相互作用,这值得我们进行仔细的考察。例如,请注意,当信度完美(等于1.00)时,预测的真值等于观测值。这并不奇怪,因为完美的信度意味着没有测量误差。反过来说,当信度为0.00(即所有方差都来自测量误差)时,预测的真值等于组均值,无论观测值的大小如何。这也不奇怪,因为在这样的情形下,以最小二乘法的视角来看,最佳的预测就是组均值。最后,当信度不完美时(大多数情形下的样子),预测的真值比观测值更接近于组均值。这是因测量误差而向均值回归的现象,第10章将要探讨的一个话题。

在下列情形下,真值会引起大家的兴趣:

1. 当我们想要匹配被试,他们来自不同的组,在组均值和所用度量的信度上存在差异。(参见 Stanley,1971:376)

2. 当我们根据不同的分数门槛,把被试分成不同的类别。(参见 Crocker & Algina,1986: 147-148)

3. 当我们在应用统计分析之前,想要校正测量误差时。例如,正如第 21 章所讨论的那样,当我们把协方差分析应用于不等值的组别时,测量误差可能会带来严重的偏差,甚至错误的结论。避免这种错误的方法之一是,采用预测的真值来进行分析,而不是采用观测值。(有关的讨论和实例,参见 Huitema,1980: 311-321)

4. 当我们想用一个标准误来设定预测的真值的信度区间时,这个论点正是下一节关于标准误的讨论时,我们所要关注的。

### 标准误和信度区间

一般来说,标准误是一个统计量的抽样分布的一个标准差。例如,一个均值的标准误是从一个既定总体随机抽样所得的很多均值的分布的标准差(参阅第 15 章有关抽样分布的讨论)。在一个既定概率下,我们可以使用标准误来设定均值的一个置信区间。(有关置信区间的讨论,参见 Hays,1988:第 6 章)

在信度的语境下,标准误存在三种不同的界定。正如斯坦利(Julian C. Stanley,1971:381)所指出的,有关它们的用法,"在文献中存在一定的混淆"(Dudek, 1979; Lord&Novick, 1968: 66-69; McHugh,1957)。我们将在这里阐述其中的两种标准误。

在当前的语境下,估值 $s_e$ 的标准误是一个既定观测值的预测真值的标准差:

$$s_e = s_x \sqrt{r_{xx}(1 - r_{xx})} \tag{5.25}$$

其中 $s_x$ 是一个既定组的度量的标准差,$r_{xx}$ 为信度系数。

对一组被试来说,假定相干的数据如下:

$$\overline{X} = 80 \quad s_x = 10 \quad r_{xx} = 0.84$$

代入公式(5.25):

$$s_e = 10\sqrt{(0.84)(1 - 0.84)} = 3.67$$

例如,如果有两个观测值 75 和 85,把它们代入公式(5.24),它们的预测的真值为:

$$(1 - 0.84)(80) + (0.84)(75) = 75.80$$

和

$$(1-0.84)(80)+(0.84)(85)=84.20$$

正如前面已经指出的那样，预测的真值比它们各自的观测值更接近于均值。

假定我们想设定这两个预测的真值的90%置信区间。在正态分布下，覆盖中间90%的部分（两侧都留下5%）所对应的 $z$ 分为1.65。因此，它们的置信区间为：

$$75.80 \pm (1.65)(3.67)$$
$$84.20 \pm (1.65)(3.67)$$

可见，对观测值为75的被试来说，最佳的预测值是75.80；约90%的真值会落在69.74(75.80－6.06)和81.86(75.80＋6.06)的区间。对得分85的人来说，最佳的预测值是84.20；约90%的真值会落在78.14至90.26的区间。我们需要作几点说明：

(1)对任何一个被试而言，真值是未知的，因此，我们无法知道它是否落在这个置信区间内。

(2)这个置信区间在预测的真值两侧对称，而不是在观测值的两侧。

(3)有些学者采用 $t$ 值（而不是 $z$ 值）来设定置信区间。当组内人数大于30时，采用 $z$ 值和 $t$ 值，实际上并没有差异。

(4)误差是正态分布的，而且，对所有的观测值来说，它们的变异度是恒定的。这是两个假定。后一个假定也称做“方差齐性”，这是第17章的讨论主题。

在设定预测的真值的置信区间时，更常见的做法是采用测量的标准误 $s_m$（参阅 Nunnally，1978：239-241）：

$$s_m = s_x \sqrt{(1-r_{xx})} \tag{5.26}$$

$s_m$ 是“当真值不变时，观测值所预期的”（Dudek，1979：335）标准差的一个估值。对比公式(5.25)和(5.26)，可见，前者在根号下多了一项（即 $r_{xx}$），这是两者的差异所在。我们需要注意两点：①$r_{xx}$ 总是一个分数，因此，估值的标准误小于测量的标准误；②当信度很高时，这两个标准误十分相似。总之，采用测量的标准误，而不是估值的标准误，将产生更大的置信区间。重要的是牢记，不管采用哪一个标准误，我们应当设定预测的真值的置信区间，而不是观

测值的。

为了和上面的计算进行对比,我们将采用相同的数据来计算测量的标准误,然后再用它来设定前面用过的、相同观测值的置信区间。

$$s_m = 10\sqrt{(1-0.84)} = 4.00$$

如前所述,测量的标准误大于估值的标准误(3.67)。

对前面所使用过的观测值(75 和 85),它们的预测的真值当然和前面的结果一样,分别是 75.80 和 84.20。使用 $s_m$,它们的 90%置信区间是:

$$75.80 \pm (1.65)(4.00)$$
$$84.20 \pm (1.65)(4.00)$$

对当前的数据来说,采用 $s_m$ 时,置信区间的规模是 13.20( =2 * 1.65 * 4.00);相对照的是,采用 $s_e$ 时,置信区间的规模是 12.11( =2 * 1.65 * 3.67)。

我们把信度系数用做一个度量的精度的一个总指数,但是,当我们的目标是评估既定分数的精度时,估值(或测量)的标准误和置信区间才会引起我们的兴趣。置信区间就像一个小贴士,它提醒我们,观测值不是没有误差的,我们应当小心谨慎地处理观测值之间的差异。

最后,如前所述,其他条件保持不变时,组间的变异度越大,信度系数越高。不过,考察标准误的公式将表明,标准误的增大将引发信度的相应增大(作为变异度增加的结果)。这等于说,信度系数可能具有较大的组间差异,但一般来说,标准误在各组间却是相似的,因此,它们是比较变异度不同的群体(组)的一个切合实际的工具。

### 低信度的逆效应

人们花费了很多精力来确定随机测量误差所带来的各种效应的方式和程度,包括特定方面(例如,参数估计、显著性检验)、设计(例如,实验、准实验)和分析流派(例如,回归分析、协方差分析)。关于测量误差对统计分析的各种效应,一个较好的综述是科克伦(Cochran, 1968)。

当我们有可能采取校正措施,或者当我们在得出结论、阐释分析结果时把低信度的效应考虑在内时,了解低信度的效应就是十

分重要的。这里我们只讨论低信度对相关系数的效应。在更复杂的情形下,低信度的逆效应不仅更复杂,而且可能“变得更致命”(Fleiss & Shrout,1977:1190)。有些复杂情形,我们将在后面的章节中加以讨论(例如,第 15 章、第 17 章和第 21 章),并提供相干的参考文献。

## 相关系数

众所周知,皮尔逊相关系数是测量两个变量之间的关联强度的最常用的度量之一。在前面的章节中,我们在效度的语境下讨论了它的应用。在本章,我们将在信度的语境下讨论它的应用。

假定我们对两个变量($X$ 和 $Y$)的相关系数感兴趣,而且我们拥有这两个变量的度量的信度估值,记为 $r_{xx}$ 和 $r_{yy}$。可以证明:

$$r_{xy} = r^*_{xy}\sqrt{r_{xx}r_{yy}} \tag{5.27}$$

其中 $r_{xy}$ 是 $X$ 和 $Y$ 之间观测到的相关系数;$r^*_{xy}$ 是 $X$ 和 $Y$ 的真值之间的相关系数。

考察公式(5.27)可见,只有当两个度量的信度完美时(即 1.00)时,$r_{xy}$ 才等于 $r^*_{xy}$。由此可见,一个或两个变量的度量的低信度将导致相关系数的一个向下的偏差或衰减。很明显,信度越低,变量间的真相关系数的估值也越低。

在第 4 章,我们注意到,旨在测量相同建构的测量工具之间的低相关系数,会让我们陷入困局,甚至让我们感到尴尬。除了效度问题之外,公式(5.27)表明,旨在测量相同建构的测量工具之间的低相关系数,至少部分来自于低信度。同理,当两个建构间的相关系数显著低于理论预期时,也是如此(参见第 4 章“跨结构分析”)。

举一个例子,让我们再次回到本章前面所评述过的米舍尔等人的研究(Mischel et al, 1974)。他们建构了一个控制点的度量,包括两个子量表,但得到的信度估值极低(0.14 和 0.20)。不出所料,所报告的两个子量表间的相关系数接近 0(“男性、女性和总样本的相关系数分别是 0.03、-0.06 和 -0.02”(Mischel et al, 1974:270)。同样不出所料的是,他们发现,子量表和其他变量间的相关系数一般都极低。

## 衰减“校正”

应用公式(5.27),我们建议对衰减进行下列“校正”:

$$r_{xy}^{*}=\frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}} \tag{5.28}$$

其中各项的定义和公式(5.27)相同。我们与其说它是衰减校正的一个公式,还不如说它是当两个变量的度量都完全可信时,相关系数的一个估值。

例如,假定 $r_{xy}=0.6$,$r_{xx}=r_{yy}=0.75$,代入公式(5.28):

$$r_{xy}^{*}=\frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}=\frac{0.6}{\sqrt{(0.75)(0.75)}}=0.8$$

可见,已知 $X$ 和 $Y$ 相关系数为 0.6,在完全可信的度量下,它们间的相关系数估计为 0.8。

我们可以在公式(5.28)的分母中,只采用一个变量(例如,$Y$)的信度估值,然后,它就衍变成一个变量的低信度校正公式。例如,这种校正可以运用于准则关联的效度研究,但却不适宜于校正预测变量的低信度(参见 Ghiselli et al,1981:290-291;Nunnally,1978:238),因为我们必须在可用的、容易出错的预测变量的分数上进行决策(例如,录取)。

正如纽纳利(Nunnally, 1978)所指出的那样,与其相信完美信度的神话,不如估计当一个变量或两个变量的度量的信度都增加一定量的时候,两个变量间的相关系数有多大,这种做法常常更有意义。方程(5.28)也适用于这类情形(参见 Nunnally,1978:238-239)。

公式(5.28)虽然比较灵活,应用也是直截了当的,但其效度受到质疑。(例如,Johnson,1950;Winne & Belfry,1982)当我们应用信度的较低估值时,我们却高估了剔除衰减后的相关系数,它甚至可能大于 1.00!不加节制地依赖剔除衰减后的相关系数,不仅可能会把我们引向一个虚幻的世界,而且可能会把我们的视线引向提高所用度量的信度的紧迫要求之外。

## 观察者间相合和信度

在各种学科和专业(例如,心理学、人类学、教育学和市场营销学)中,把观察者作为一种测量方法都很普遍。我们将在第 6 章中讨论观察,并对信度作评述,给出相关的文献。这里我们仅指出,区分“观察者间相合”和“观察者间信度”(有些研究者称之为“评

估者相合”和“评估者信度”）十分重要，而且，我们有很多种方法来测量它们中的每一个。

## 概括度理论

估计信度的传统方法所面临的困难之一是，尽管我们知道，测量误差可能具有不同的来源，但在估计过程中它们纠缠在一起。在完全的重新表述之后，克伦巴赫等人（参阅 Cronbach et al，1972）发展了概括度理论。和真值的古典测量理论不同，在概括度理论中，“研究者把观测值或者它的一种函数当做是总体值，即他从样本概括到总体。‘信度’问题因此转换成概括的精度（或概括度）问题”（Cronbach et al，1972：15）。

采用复杂的方差分析设计，概括度理论让我们能够同时鉴别和区分不同的误差源（例如，被试、场合、评估者、题器和时间）。运用概括度理论的必要前提是洞悉方差分析。而且，大家必须明白，我们想探讨的误差源越多，设计就越复杂，执行就越艰难。这可能是概括度理论较少得到应用的原因所在。

如果大家有兴趣学习概括度理论，我们建议大家不要从克伦巴赫等人的著作开始，那很复杂，而要从一些导论开始。或许布伦南（Robert L. Brennan，1983）的导论是最好的，它还有伴随的计算机程序，分析按照概括度理论所进行的设计。（同时参见 Crocker & Algina，1986：第 8 章；Feldt & Brennan，1989；Shavelson et al，1986；Webb et al，1988）

**练　习**

1. 一个测量的误差方差和总方差分别是 11.58 和 73.62,它的信度估值有多大?

2. 一个度量由 10 个题器组成,它的信度估值是 0.50。当它的长度分别增加到:

(a)20 个题器,

(b)30 个题器,

它的信度期望各自有多大?

3. 一个测试由 100 个题器构成,它的信度估值为 0.92。假定我们想要精简这个测试,让精简版测试的信度估值等于 0.82。它的题器数是多少?

4. 一个测试由 15 个题器构成,题器的方差和等于 8.36,总方差等于 23.61。

(a)α 信度的估值是多少?

(b)该测试的方差中,有多大比例源自误差?

5. 下列数据是 10 个被访者对 8 个二项题器的答案(即 1 为正确答案,0 为错误答案)。

1 1 1 1 1 1 1 0
1 1 1 1 1 0 1 1
1 0 1 0 0 0 0 0
1 0 1 0 0 0 1 1
0 0 1 0 0 0 0 0
1 0 1 1 1 1 0 1
1 0 0 0 1 0 0 0
1 0 0 1 0 0 1 0
0 1 0 1 0 1 0 0
0 0 1 1 1 0 1 0

求:

(a)题器均值?

(b)题器方差?

(c)总分的均值? 证明它等于题器均值之和。

(d)总分的方差?

(e)用 α 系数所得的该测试的信度?

如果你可以使用信度的计算机程序(例如,SPSS),我们建议您运行这个示例,并和您的手算结果进行比较(我们已经在本章给出了 SPSS 命令语句的示例)。

6. 假定我们从一个相对较大的样本中估计出：

$$\bar{X} = 62 \quad s_x = 7.63 \quad r_{xx} = 0.75$$

（a）一个被试的观测值是65，请估计他的预测真值。

（b）估值的标准误是多少？

（c）测量的标准误是多少？

（d）利用估值的标准误，计算预测真值的68%的置信区间。

（e）利用测量的标准误，计算预测真值的68%的置信区间。

7. $X$ 和 $Y$ 的相关系数是0.62，它们的信度估值分别是0.82和0.73。假定这两个变量的度量完全可信，求 $X$ 和 $Y$ 间的相关系数。

**参考答案**

1. 0.84

2. （a）0.67——参见公式（5.15）

（b）0.75

3. 40个题器——参见公式（5.16）

4. （a）0.69

（b）0.31

5. （a）0.7　0.3　0.7　0.6　0.5　0.3　0.5　0.3

（b）0.21　0.21　0.21　0.24　0.25　0.21　0.25　0.21

（c）3.9 = 0.7 + 0.3 + 0.7 + 0.6 + 0.5 + 0.3 + 0.5 + 0.3

（d）4.09

（e）0.64

6. （a）64.25

（b）3.30

（c）3.82

（d）60.95　-67.55

（e）60.43　-68.07

7. $r_{xy}^* = 0.80$——参见公式（5.28）

# 6

# 社会行为研究中的几种测量方法

许多书籍和众多论文都致力于考察社会行为研究中的测量方法(例如,多项选择题测试、评分量表、投射技术、访谈、观察方法),甚至考察一个方法中的具体方面(例如,题器特征、测试等值、常模、应答类型)。有些方法已经得到发展,并主要应用在特定的研究领域中(例如,态度、绩效、心理能力、人格、爱好)。有一些方法(一般称为“测量模型”)关注于区分人与人,而另一些方法(一般称为“量表模型”)关注于区分刺激与刺激。下列书籍是探讨测量或量表(或两者)问题的浩瀚文献中的极少一部分:库姆斯(Coombs, 1964),考克森(Coxon, 1982),爱德华(Edwards, 1957b),吉尔福德(Guilford, 1954),克鲁斯卡尔和威什(Kruskal & Wish, 1978),马拉内尔(Maranell, 1974b),麦基弗和卡迈恩斯(McIver & Carmines, 1981),纽纳利(Nunnally, 1978),托格森(Torgerson, 1958)以及范德维(van der Ven, 1980)。

在一章的篇幅内,我们显然不可能对如此广阔的领域作出一个极为粗略的综述。我们的目的仅局限于介绍一些方法,探讨一些和方法关联的问题。入选本章的各种方法的唯一理由是,它们是最流行的方法之一,同时也是因为它们可以服务于各种各样的目的。

首先,我们将讲解一般意义上的评分量表,然后是累加评分量表和语义微分量表。随后,我们将讲解访谈,强调结构效应和访问员效应。任务效应和被访者效应的一些问题,则分章节进行了讨论,因为它们的效应并不局限于访谈。讨论观察的一节是本章的

结尾。①

## 评分量表

“无处不在”(Dawes, 1972:93)和充满诱惑的评分量表已经有很长历史了,它们的应用至少可以追溯到“公元前150年,希巴克斯(Hipparchus)用6点量表来判断星星的亮度”(Lodge, 1981:5)。

您不仅见到过不同形式的评分量表,而且您还有可能偶然回答过评分量表。我们相信,这样的假定并不会太离谱。也许您还使用过评分量表来取得别人的意见。无论如何,您肯定知道,我们使用评分量表来量化对自己、他人、事物和场景的评价、印象、判断和知觉(列举一个常见的领域)等。

评分量表如此流行的原因,很可能是因为它们的编制和管理都相对比较容易,而且它们似乎可以应用于测量所有能够想象得到的事物。尽管评分量表具有很多种形式,但它们具有一个共同点:它们都要求被访者参照一个评级(例如,对一个态度陈述的赞同程度,一个行为的出现频率,一个产品的质量)来指明自己的立场(最宽泛的意义上)。

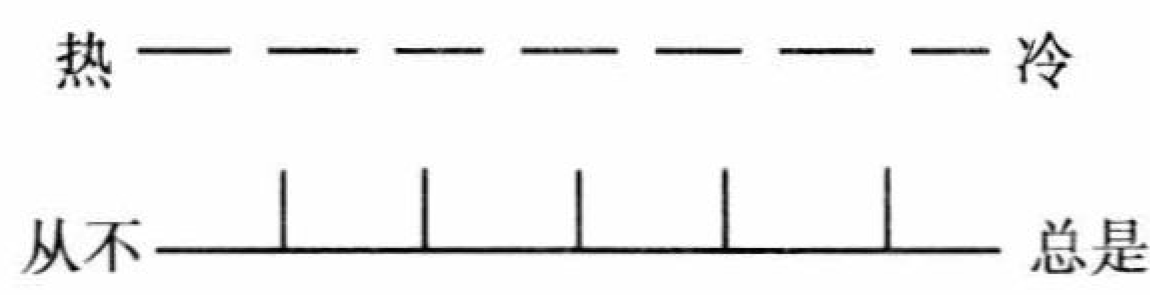
### 量表格式

不同格式的评分量表都有应用。下面我们给出一些流行格式的例子,并作简要的评述。(详细讨论,参见 Aaker & Day, 1983; Dawes, 1972; Dawes & Smith, 1985; Gable, 1986; Guilford, 1954; Lemon, 1973; Lin, 1976; Nunnally, 1978; Saal et al, 1980)

最流行的评分量表可能是图式评分量表,它和被访者所熟悉的测量策略(例如,用尺量长度,用温度计测温度)有相似之处。一个图式评分量表一条直线构成,两个端点标记为准星(例如,懒—勤、热—冷、完全赞同—很不赞同、从不—总是)。我们要求被访者要在这条直线上标出一个点,对应或反映他们的立场。这条直线常常是虚线或分割成不同的线段,如下所示:

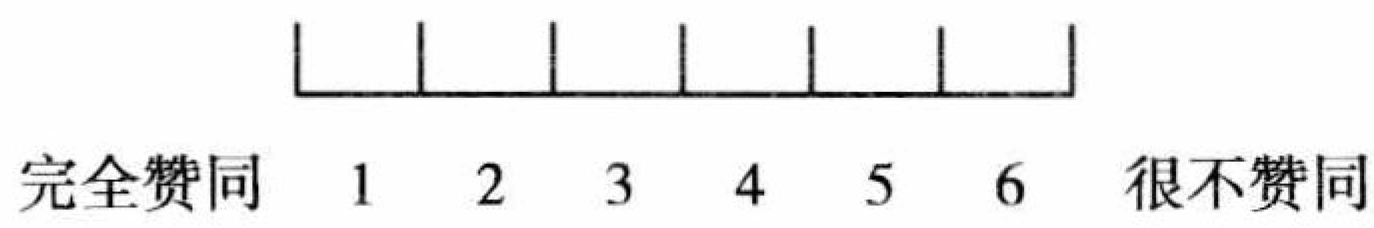
---

① 在阅读本章时,特别是阅读讨论访问员效应、任务效应和被访者效应的章节时,浏览第11章的相关章节,可能会有所帮助,因为它们在很多方面是对本章的补充。




被访者虽然可以把标记画在一个线段的任何位置上,但通常我们等同处理一个线段内的标记。因此,我们把上面的量表看做是7点或7级(在后面的章节我们将讨论最优点数或级数的问题)量表。相应地,我们将一个量表上的答案编码为1到7分,1表示一个端点(例如,从不),7表示另一个端点(例如,总是)。

在上面的例子中,数字是隐含的。在下面的例子中,数字则显示在量表上:




我们不仅可以描述和界定量表的两个端点,也可以描述和界定量表的一些或全部中间点,标上数字或不标上数字。例如:



| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 从不 | 很少 | 经常 | 总是 |

在很多情形下,为了方便被访者,我们会给他们提供一组事先界定的选项,让他们就各种陈述、行为、特质等,指明自己的立场。例如,一组态度题器中的每一个题器前面,可能会有一条短线,要求被访者在上面记录自己赞同或不赞同的程度。下面是这种应答选项的三种不同格式的示例,包括应答选项的定义:

| (a) | (b) | (c) |
|---|---|---|
| +3:十分赞同 | 6:十分赞同 | AVS:十分赞同 |
| +2:比较赞同 | 5:比较赞同 | AS:比较赞同 |
| +1:赞同 | 4:赞同 | A:赞同 |
| -1:不赞同 | 3:不赞同 | D:不赞同 |
| -2:较不赞同 | 2:较不赞同 | DS:较不赞同 |
| -3:很不赞同 | 1:很不赞同 | DVS:很不赞同 |

在这些(或其他可能的)格式中作选择,取决于很多因素。例如,当被访者比较练达时,我们一般倾向于使用格式(a)。把赞同与正号相联系、不赞同和负号相联系,这种倾向有助于保持使用这

些事先界定的类别时的连贯性。但是,对有些被访者(如文盲、儿童),格式(a)就显得不恰当。

## 评分量表的一些共同特征

评分量表除了具有独特的特征和格式之外,还具有一个主要特征,即从事评分的个人担当测量工具的角色。由此可推,评分的效度和信度的基础是“一个假定,即人类观察者是定量观察的一个良好的工具……拥有一定程度的精度和一定程度的客观性”(Guilford, 1954:278)。这样的话,毫无疑问,评分的效度和信度会有很大的变化,它取决于谁来做评分、在何种参照系下、为了什么目的、在何种情境下,等等。这让一些研究者(如 Oppenheim, 1966)担心,甚至怀疑评分量表的有用性。

下面,我们将简要评述一下使用评分量表时所遭遇的一些主要问题。

### 评分与知觉

评分量表容易遭到误用和误解,主要的原因很可能是由于它们源自一个事实:它们反映了一个知觉过程。研究广义知觉和人类知觉的文献中(参见 Markus & Zajonc, 1985; Schneider et al, 1979; Warr & Knapper, 1968)充斥着知觉者的态度、价值观、动机等影响知觉的例子。因此,毫不奇怪的是,评分常常透露的是评分者本人,而不是反映他所评估的对象。

在各种来源的评分者偏差和选项集面前,评分十分脆弱。这个现象已经有很多的文献记录(Guilford, 1954; Landy & Farr, 1980; Saal et al, 1980)。最常见的偏差之一是晕轮效应,即评分者对被评估对象的整体印象会扭曲他们对其各个方面的评分,产生恒定的误差。另一种偏差是有些评分者总是倾向于打高分或打低分(宽严误差)。有些评分者倾向于避免选择极端的类别,而集中于选择量表中点附近的类别(趋中误差)。为了让评分者偏差最小化,我们可以采用各种策略,包括在应用具体量表时进行培训,对所采用的参照物和量表进行清晰的界定。我们仅就后者作一点评述。

### 定　义

我们在其他章节(比如,第 3、4、8 章)讨论了有关定义的一般

问题。这里我们强调,对有待评估的参照物和所用评分量表的清晰界定是绝对必要的。必要的详略程度则随特定的参照物、量表类型、被访者、情境等的变化而变化。

令人遗憾的是,我们提供给评分者的评分量表常常没有任何界定。极端的情形是,我们要求他们给各方面一个总评(例如,一个员工多么“好”,一个教授如何“多产”,一个领导多么“民主”),却没有对相关的术语进行定义。在缺乏界定的时候,评分者显然只能诉诸自己的定义或概念。毫无悬念,在这种情形下,评分的效度和信度倾向于很低。

甚至当我们要求评估者就具体的属性或方面给予评分时,也常常是没有定义或定义模糊,典型的示例是卡特总统的白宫主管乔丹(Hamilton Jordan)所制订的《员工评估表》,它由 30 个题器构成,是让内阁官员和白宫员工用来评估自己的下属。下面是一些例子:(a)受评者的自信程度如何(从“自疑”到“自大”,“自信”位于中间)?(b)他/她的稳定程度如何(从“摇摆不定”到“坚定不移”)?(c)他/她的消息面如何(从“窄”到“宽”)?这个评估表的拷贝,参见《纽约时报》1979 年 7 月 19 日 A16 版。

乔丹的方法与广泛应用于私营产业和学术界的方法并没有不同。下列题器是从学生评价教授的各种量表中随机抽取的:(a)原创—传统,(b)创造—常规,(c)善于接纳新思想,(d)对教学感兴趣,(e)关心学生,(f)讲解能力和(g)评价学生时一视同仁。(这些摘录题器的量表和其他量表,参见 Elmore & LaPointe, 1975; Marsh, 1982; Sockloff, 年份不详)

类别定义。已经证明(参见 Goocher,1965,1969; Hakel,1968; Simpson, 1944),评分量表所采用的类别(例如,时常、经常、偶尔、有时),表面上看起来并没有歧义;但不同的评分者会给予不同的阐释;对不同的情境下的同一个评分者,它们的意义也不相同。

界定的一个更具体的关切(常常称做“定准”)和两个问题有关:我们是否应当界定所有或部分类别?什么是界定的最佳模式(即端点、中点)。涉及这些问题的研究所得到的结论并不一致。到目前为止,我们还不清楚何种程度的类别界定(例如,全类别界定、部分类别界定、端点类别界定)是最优的。(有关综述,参见 Dixon et al, 1984)正如兰迪和法尔(Landy & Farr, 1980:88)所指出的:

> 准星的类型和数量的重要性，很可能和维度界定的恰当性关联在一起。当有待评估的维度缺乏恰当的界定时，评分者就必须依靠准星来决定这个量表的意义。

## 评分的复合

我们常常把几个评分量表上的分数合并成一个复合分，以方便作出有关受评者的决策（例如，提升谁、解雇谁、奖励谁终身教职）。毋庸赘言，每个评分应当诉诸相同的准则或相同准则的几个方面，否则的话，这个复合分注定是没有意义的或误导人的。不过，即使每个评分都诉诸相同的准则，我们仍然会面对一个极端艰难的问题：在取得总分的过程中，如何给每个评分赋予权重？

把所有评分相加或求均值，也就是给每个评分相同的权重，这很可能是最经常采用的方法，但常常也是不恰当的做法。在评价员工时，一个雇主赋予按时上班、听从指挥、保持工作间整洁、精准、有效率等相同的权重，我们可以有把握地说，这种情形十分罕见。每个方面的相对重要性将取决于这个准则的具体定义（假定我们已经做出过尝试），而这个定义又取决于具体的岗位和具体的场景。

没有给出理据或给出极少的理据就合并评分，甚至稍加审视就可以看出它是不恰当的，不幸的是，这样的例子不胜枚举。表6.1是评分制的一幅漫画，作者设计它的目的，是想让它成为过分滥用评分的一剂解药。但遗憾的是，我们并不知道这个作者的身份。

**表6.1　员工绩效评估表**

| 绩效领域 | 绩效程度 | | | | |
|---|---|---|---|---|---|
| | 远超岗位要求 | 超过岗位要求 | 符合岗位要求 | 有待提高 | 不及最低要求 |
| 工作质量 | 单跳过高楼 | 助跑过高楼 | 撑竿过矮楼 | 撞上大楼 | 不识大楼 |
| 快捷性 | 比出膛的子弹快 | 和出膛的子弹一样快 | 您会相信一颗慢速子弹吗？ | 经常哑火 | 开枪自伤 |
| 适应性 | 水上行走 | 顶住压力把头伸出水面 | 以水洗面 | 呛水 | 遇事就溺尿 |
| 沟通 | 与上帝交流 | 与天使交流 | 与自己交流 | 与自己论证 | 败下阵来 |

## 累加评分量表

在前面的章节,我们已经指出,就效度和信度而言,单题器的度量常常存在不足。因此,我们才经常诉诸多题器量表,有关成绩、才能、人格和态度的度量都是例子。

一般来说,一个累加评分量表是指一个包含多个题器的量表,我们把这些题器上的得分加总,求得一个分数。这种量表的最流行的形式之一,当前仍在使用的,可能是李克特量表,它取名于李克特(Likert,1932),是他首创了这种建构量表的方法。

建构李克特量表的第一阶段是建设一个题器库。我们可以自己设计题器,同时(或者)从各种来源(例如,文献、大众媒体、日常话语)中收集有关参照物的各种表述,来完成这项任务。为了让选项集最小化(例如,不论题器内容如何,一律表示赞同①),我们应收集有关参照物的(大约相同数量的)正面和负面表述。

我们首先应把褒义(正面)题器或贬义(负面)题器的计分倒过来,让赞同正面题器和不赞同负面题器,得分相同。例如,如果"十分赞同"一个正面题器得 6 分,那么,"很不赞同"一个负面题器也应该得 6 分。然后,我们才能把各种题器的得分加起来,得到一个总分。

### 题器分析

我们把题器库应用到一个甄别样本上,它的被试构成和该量表想要测量的人群相似。这些被试回答量表上的每一个题器,并指出自己赞同/不赞同的程度。最初,李克特使用 5 点量表(即"很赞同""较赞同""说不清""较不赞同""很不赞同"),后来在这种量表中,人们使用过不同数量的选项(包含"说不清"或删除"说不清"选项)(参阅本章稍后的"任务效应")。

我们对甄别样本的答案进行题器分析,以便决定每个题器的恰当性,并选择"最佳"题器进入量表中。李克特提出了两种选择题器的方法:(a)选择能够成功区分"高分组"和"低分组"的题器②(有关题器鉴别力的检验,详见 Edwards, 1957b:第 6 章)或者(b)

---

① 参阅本章后面的"应答风格"。

② 以他们的总分数为基础,大约选上下 25% 作为每一组的数据。

选择那些和总分相关系数高的题器。一个题器是总分的一部分，因此，题器和总分间的相关在一定程度上是通胀的。有观点认为，我们需要校正这样的相关系数。在过去，我们通过采用一个估计公式来进行校正（Nunaally, 1978：281）。现在，计算机程序（例如，SPSS 的 RELIABILITY 程序）可以在计算题器和总分的相关系数时，把特定题器排除在总分的计算之外。当题器数相对较少时，这种校正尤其重要。

## 维　度

当我们认为一个总分反映一个建构时，所有和建构验证有关的问题（参见第 4 章）都是直接相干的。维度问题就是一个很切题的问题。正如我们在第 4 章所指出的，题器总分的效度是由它们触及相同维度所预测的。而且，就确定一个量表的维度而言，前一节所描述的题器分析，作用就十分有限了。

因子分析是研究一组题器间的关系背后的维度的一种方法（直观介绍参见第 4 章，详细讨论和示例分析参见第 22 章和第 23 章）。在当前的语境下，在对甄别样本进行测量的题器库中（参见前一节），选择题器的最佳方法很可能就是对它们的相关矩阵进行因子分析。在同一个因子上具有高负荷，在其他因子上具有低负荷的题器，我们可以认为它是测量当前维度的一个量表的较佳候选者。采用这个方法，我们可以生成和验证一个量表，它旨在测量单维建构或多维建构。在前一种情形下，我们选择在单个因子上高负荷的题器进入这个量表。在后一种情形下，我们选择在不同因子上高负荷的题器进入子量表，每个子量表形成当前维度的总分。

## 题器计分

最初，在正态离差的基础上，李克特提出了一种相对比较复杂的题器计分方式。不过，已经证明，给每一个类别赋值一个整数（即 1 =“很不赞同”，2 =“较不赞同”，等等）是更简单的方法，但却可以产生和前一种费力的方法十分相似的结果。

## 总分及其阐释

尽管我们可以给每个题器赋予不同的权重（例如，在计算因子分时，采用从因子分析所取得的权重），但就大多数目的而言，已经

证明，权重为1（即只把每个答案加总，让每个题器具有相同的权重）就可以带来令人满意的结果。

我们可以把总分看做是单个题器之和，但更有用处的表述是把它看做一个均值，即总分除以题器数。例如，假定一个度量由20个题器构成，选项从1＝"很不赞同"到6＝"十分赞同"。再假定一个人的总分是96。若想了解这样一个总分的意义是很困难的。但如果把它除以20（题器数）得到一个4.8的分数，在上述量程上，我们就比较容易阐释它。

但我们应当注意到，无论表达为总分还是均值，累加评分量表上的分数，本质上都是相对的。因此，我们可以把一个人的位置和他（或她）的群体或其他常模相比，我们也可以对不同群体进行相互比较。

## 语义微分

概念、物体或个人等刺激会引起多种反应。我们可以采用"语义微分（SD）"技术来评估这些反应。它最初由奥斯古德（Osgood，1952）所提出，随后由奥斯古德等人（Osgood et al，1957）所发展。SD[①] 旨在评估各种概念的内涵或喻意。它的功能基础是两个基本概念：（a）概念的差别在于它们所传达或所引发的意义不同（因而我们才可能将它们加以"区分"），（b）用相对很少的维度，我们就可以把握大多数概念的意义（我们将在后面加以讨论）。

我们在一组两极形容词量表上来给概念评分。有时，量表结构是事先决定的；除此之外，我们使用SD一般带有双重目的：（a）通过考察量表间的关系来研究各种量表的意义，（b）评估概念的意义或概念之间的差异。

表6.2列出了SD中最常用的格式，我们将在9个两极量表上对"计算机"概念进行评分。[②] 我们设计这些7点量表，既用来评估方向（例如，是好还是坏？），也用来评估强度（例如，多好或多坏？）。很明显，SD是一种特殊类型的评分量表。因此，先前讨论过的有关

---

① 为方便计，我们采用SD一说，但这并不意味着我们暗示它是一种度量。相反，它是一种具有广泛应用的技术。

② 作为示例，我们在这里只给出了一个概念。我们将在后面讨论与选择概念和量表相关的问题。

评分量表的一般问题或关切（例如，定义、复合分）也适用于SD。

**表6.2 语义微分的概念和量表示例**

| 计算机 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 好 | —— | —— | —— | —— | —— | —— | —— | 坏 |
| 慢 | —— | —— | —— | —— | —— | —— | —— | 快 |
| 丑 | —— | —— | —— | —— | —— | —— | —— | 美 |
| 主动 | —— | —— | —— | —— | —— | —— | —— | 被动 |
| 大 | —— | —— | —— | —— | —— | —— | —— | 小 |
| 弱 | —— | —— | —— | —— | —— | —— | —— | 强 |
| 贵重 | —— | —— | —— | —— | —— | —— | —— | 廉价 |
| 无力 | —— | —— | —— | —— | —— | —— | —— | 有力 |
| 锋利 | —— | —— | —— | —— | —— | —— | —— | 迟钝 |

考察表6.2中的量表，请注意，尽管每一个量表都具有细微的差别，但它们并不表示意义的9个不同维度。我们并不需要花费多少力气，就可以把它们分成子集，每个子集反映意义的一个共同维度。例如，好—坏和美—丑看起来拥有共同的意义，快—慢和主动—被动也一样。

事实上，对上述量表之间关系的因子分析（Osgood et al.，1957），总是重复形成三个主维度：(a)评价(E)，它通常是最主要的一个维度，是指对反应的舒服程度，即"人类思维中的态度变量"；(b)性能(P)，它是"指能量及其相关的事情，大小、重量、韧性等"；以及(c)活动(A)，它包括"迅速、兴奋、温暖、激动等"（同上：72-73）。

参照表6.2中的各个量表：好—坏、美—丑、贵重—廉价反映的是评价维度；大—小、弱—强、无力—有力反映的是性能维度；快—慢、主动—被动、锋利—迟钝反映的是活动维度。暂时假定这三个维度具有效度，把构成特定维度的每个量表的答案加起来，我们就会求得每一个维度上的得分。

为了让选项集最小化，我们常常以两极平衡的方式排列量表，让一极的形容词（例如，正面）和另一极相平衡（参见表6.2，其中"好"和"丑"都在左边）。因此，正如我们在累加评分量表所解释

的那样,在相加之前,相关量表的计分必须倒过来。然后,我们就可以使用各种单变量或多变量分析,来评估两个或多个概念(以及两组或多组概念)之间的差异。

在使用 SD 时,我们会碰到一些问题和关切(参见下文),但它还是得以延续下来,部分原因在于,针对不同的总体,各种研究所获得的 EPA 维度具有相对的一致性,它们解释掉评分中的大部分协方差。[①]

SD 具有较大的普及率,因为它容易管理,对被试应答的要求也比较简单,而且它适用于范围广泛的话题。但正是这些方面带来了 SD 应用中的大部分混淆和误解,它们围绕着对数据的分析以及对结果的阐释。

和累加评分量表的情形一样,很多研究者似乎认为,如果把各种两极评分量表和一组概念连起来,他们事实上就在应用 SD。同时,我们还应当注意到,有些量表本质上并不是真正的两极量表,但却被使用于所谓的 SD 应用中。

最初,SD 的设计目的是为了弄清一组概念的意义。因此,人们有意使用了各种各样的概念。后来,很多(如果不是大多数)应用聚焦于特定概念域(例如,族群、自我的某些方面)的意义,或者(以及)人们对它们的态度。

对 SD 的更深入讲解超过了本书的范围。对 SD 的一般阐述,对它的假定、在应用过程中常常碰到的问题的讨论,请参阅:Bynner & Coxhead, 1979;Heise, 1969b, 1970;Maguire, 1973;Mann et al, 1979;Mayerberg & Bean, 1978;Miron, 1972;Miron & Osgood, 1966;Osgood et al, 1957;Snider & Osgood, 1969。

在本章余下的篇幅中,我们将探讨一些和选择量表和概念、概念—量表的互动、分析方法有关的问题。

## 选择量表

最初,奥斯古德及其同事构建了 50 个 7 点评分量表,他们采用常用的一对两极形容词(例如,好—坏,大—小,硬—软,甜—酸,

---

① 在跨文化的研究中,这三个维度也得到重复(例如,Osgood et al. ,1975; Snider & Osgood, 1969:第五部分)。斯奈德尔和奥斯古德编辑的集子里,有 SD 在不同背景和研究领域中(例如,社会心理、人格、美学和广告)应用研究的很好示例(Snider & Osgood, 1969)。

强—弱,干净—肮脏,高—低,冷静—冲动)定准量表的端点。他们要求被试用这50个量表去评估20个不同概念(参见下文)。研究的中心任务是发现量表背后的维度;因此,他们把每个被试和概念上的分数加总,形成了一个所有量表间的50×50的相关系数矩阵(我们将在后面讨论这种和其他建构相关系数矩阵的方法,并讨论各种分析方法)。

对50×50相关系数矩阵的因子分析,得到了前面所提到的三个主要因子(即EPA),有些量表在一定程度上是一个特定维度的单纯度量(例如,好—坏与E;有力—无力与P;快—慢与A)。奥斯古德等人(Osgood et al, 1957)报告了这50个量表的因子负荷,同时他们还对其他研究作了一个小结;在所采用的量表、概念和被试上,这些研究存在一定的差异。

在后来的研究中,在每个维度上,我们经常使用3~4个代表性的量表,一般来说,我们可以取得具有恰当信度的因子分。奥斯古德等人的主要目的是研究量表的维度,因此,相比选择概念而言,选择量表必然更结构化、更严谨。

选择量表的准则包括因子构成和概念的相干度以及语义稳定性。为了取得EPA的各个子量表的分数,我们可以从奥斯古德等人最初的50个量表中(或者是从其他列中表)选取一个子集,但这需要一个假定,即这些量表的确能够反映这三个维度。尽管这三个维度具有相对的持久性和稳定性,但我们仍然有必要考察一个既定研究所采用的量表和概念组合的因子结构。如果我们不能确信EPA结构是适用的,不能确信假定测量一个维度的量表实际上的确如此,那么,我们就会面临很多困难(我们将在后面讨论一些困难)。在概念域狭义界定的研究中,"通常的相关系数结构已经遭到破坏"(Kahneman, 1963: 554),情形尤其如此。

这并不是暗示,在选择量表时,我们不必参考以前利用SD所进行的研究。但我们想要强调的是,大家最好能够遵守梅尔伯格和比恩的两个"不要"(Mayerberg & Bean, 1978:479):

> ①不要在以前研究的基础上假定量表的意义;②假定反映相同意义维度的量表,如果没有证据表明它们的确如此,不要把它们的分数加总。

## 选择概念

如前所述,在SD的原初概念中,我们把重点放在了量表的意义上。因此,选择概念并不像选择量表那样显得结构分明。选择概念的准则也不过是概念的尽可能多样化和被试对概念的熟悉而已。奥斯古德等人(Osgood et al,1957: 34)写道:"在此基础上,实验员简单选择了下列20个概念:女士、巨石、原罪、父亲、湖泊、交响乐、俄罗斯、羽毛、我、火、婴儿、骗子、上帝、爱国者、龙卷风、剑、母亲、雕像、警察、美国。"

我们已经指出,在大多数SD的应用中,人们感兴趣的是特定的实际问题。很显然,在这样的应用中,概念应当对问题域具有"代表性",而且为被试所熟悉,但它们不必局限于让所有被试在所有量表上产生相同(或接近相同的)反应的概念。[①] 因为有限的变异度将严重制约相关系数(参见第3章"范围限制"一节)。在这种情形下,如果对相关系数进行因子分析,我们很有可能发现不了任何结构。

选择概念(和量表)的其他相关问题,我们将在下面讲解(即"概念—量表互作"一节)。这里,我们继续引用梅尔伯格和比恩(Mayerberg & Bean,1978:479)余下的"不要":

> ③除非有证据表明对不同概念的反应高度相似,否则不要在一个概念域中把各种概念的得分相加;④不要把反映不同概念域的概念得分相加。

## 概念—量表互作

两个量表可能拥有意义的一个共同成分,独立于被评估的概念本身之外,因而让它们具有相互联系。另外,概念(刺激本身)可能会在一定程度上决定量表间的关系(Bynner & Coxhead, 1979)。换言之,量表可能会有区别地关联到概念上,并且(或者),概念也可能会"引发形容词的语义漂移"(Heise, 1969b:416),并导致相同量表间的关系差异,视这些量表所应用的概念而定。这个现象所

---

① 有些被评估的概念接近于EPA维度的端点,有关例子(正面评价:家庭、教堂、真理;负面评价:蜗牛、石头、睡觉),参见Heise(1970)。

造成的一个结果(称为“概念—量表互作”)是,随着所研究的概念不同,我们可能会得到不同的因子集(性质、数量不同)。当我们只使用一个或几个概念时,这种结果更容易发生,而在这些情形下,因子分析“便和风险结缘”(Heise, 1969b:421)。

概念—量表互作绝不是一个新鲜概念。事实上,在开发 SD 的早期阶段的研究基础上,奥斯古德等人(Osgood et al, 1957:187)得出总结:“很明显,存在很高程度的概念—量表互作;随着被评估的概念不同,量表的意义以及它们和其他量表的关系都会发生相当大的变化。”他们(Osgood et al, 1957:188)接着说道:“显然,在建构广义的语义测量工具时,这些结果提出了严重的实践问题……归根到底,我们已经证明,有必要对评估的每一类概念,分别建构测量工具。”

不幸的是,后人并没有重视奥斯古德等人的忠告。概念—量表互作的大量证据对它们所带来的各种问题的详述,也没有遭遇更好的命运。(参见 Bynner & Coxhead,1979; Heise, 1969b, 1970; Kubiniec & Farr, 1971; Maguire, 1973; Mann et al, 1979; Mayerberg & Bean, 1978; Miron, 1972)

有些概念—量表互作可能是方法上的瑕疵,可以设想,我们可以减少甚至完全消除它们(例如,选择恰当的量表、使用恰当的分析单位,参见 Heise, 1969b),但真正的例子还是不胜枚举。随着被评估的概念不同,量表间的相关系数也可能不同,考虑到这个事实,可以推论,因子结构也可能在一定程度上有所不同。因此,如果我们依赖一个公认的因子结构来推算 SD 维度的分数,那么,最好的结果是得到一个(概念间的)不相干的比较,最坏的结果是得到一个错误的比较。

在大多数 SD 应用中,评估一下潜在的概念—量表互作,这种可能性甚至都没有得到考虑。人们事先就接纳了 EPA 结构,并据此给所有概念计分。甚至在进行因子分析的时候,由于生成量表间相关系数矩阵的方式,甄别概念—量表互作的可能性也常常被排除在外(有关解释,参见下文)。

总而言之,我们应当利用适合于侦测概念—量表互作出现的方法,考察每一个新概念域的结构,这是基本的要求。

### 分析 SD 数据

分析 SD 数据是一项比较复杂的任务。我们首先讨论造成这

种复杂性的主要因素,然后再提纲挈领地讲解和评述一些分析方法。

如前所述,设计 SD 的基本目的有两个:(a)考察量表的意义结构,(b)利用这种结构来评估概念的微分意义。[①] 和更常规的数据类型不同的是,SD 数据包含三个维度或模态:概念、量表和个人。[②] 我们通常采用因子分析对量表结构进行探索。目前,大多数因子分析方法仅适用于两模态的数据(即被试和变量,参见第 22、23 章)。为了把传统的因子分析法应用到 SD 数据上,某种意义上,我们需要把三模态的结构分拆成两模态的结构。[③] 依据伯恩纳和考克斯黑德的研究(Bynner & Coxhead,1979),我们将讲解完成这种分拆的五种方法。[④] 如下所述,它们的差别在于关注不同的变异源。因此,采用不同策略分拆的数据,应用因子分析所得到的结果可能会显著不同。

(1)个体 × 量表。马圭尔(Maguire, 1973),迈诺尔和奥斯古德(Miron & Osgood,1966)将这种方法称为"总和法",它计算每一个人在每一个量表的所有概念上的平均分,然后得到两模态的数据。然后,它从"个体—量表"数据矩阵,计算量表间的相关系数矩阵。实际上,它忽略了概念间的差异,把概念模态看做是最不重要的一个。"在奥斯古德的原初语境下,即通过大量概念来寻求和概念无涉的意义成分,这个方法看起来是恰当的。"(Bynner & Coxhead, 1979:376)因为它沿着概念分拆数据,因此,研究概念—量表互作的尝试就被排除在外了。

(2)概念 × 量表。与前一种方法相似,这也是一种加总方法,它计算每一个概念—量表组合下的所有个体得分的均值。它计算"概念—量表"矩阵,求得量表间的相关系数,然后再对它进行因子分析。显然,它把个体看做是最不重要的模态,因而忽略了个体间的差异。当我们的主要关注点是群体知觉时,这个方法可能是恰

---

① 考察概念差异必然取决于量表结构,因此,我们把讲解限定在考察后者的分析方法上。

② 我们可以把 SD 数据中的变化归因于三种模态及其互作。对 SD 数据背后的线性模型的全面讲解,参见马圭尔(Maguire, 1973),伯恩纳和考克斯黑德(Bynner & Coxhead, 1979)。卡恩曼提出了一个更加限定的模型(Kahnemann , 1963)。

③ Tucker(1966)开发的三模态因子分析可以直接处理 SD 数据,但这类分析的应用十分罕见。

④ Bynner 和 Coxhead 在 375 页上所制作的图示,可能有助于厘清这五种方法。

当的。

请注意,在简化的数据矩阵中,概念构成矩阵的行,这和大多数因子分析中的被试(个案)一样。因此,为了得到稳健的解,我们就需要相对数量较多的概念。伯恩纳和考克斯黑德(Bynner & Coxhead,1979)建议,30 个概念是最低限;但海斯(Heise,1969b:419)却主张“40 个概念是一个较合理的底线”。

(3)每个概念的个体×量表。它分别处理每一个概念,先计算每一个“个体—量表”矩阵,再求得量表间的相关系数。也就是说,概念有多少,“个体—量表”矩阵就有多少。马圭尔(Maguire,1973),迈诺尔和奥斯古德(Miron & Osgood,1966)建议,对从单个矩阵求得的相关系数,先求均值,再对求得的“个体—量表”矩阵进行因子分析。但这样做就等于忽略了概念—量表互作问题。因此,我们建议,首先考察每一个相关矩阵,如果量表间的相关系数在不同概念之间存在显著差异,则表明存在概念—量表互作。如果不存在概念—量表互作的话,我们可以求相关系数的均值,并对所得的矩阵进行因子分析;如果存在概念—量表互作的话,其他分析方法可能更恰当(例如,对每个概念分别进行因子分析)。

(4)每个个体的概念×量表。在前一个方法中,我们分别处理每个概念,这里我们分别处理每个个体。也就是说,对每一个个体,我们计算“概念—量表”数据矩阵,然后再求得量表间相关系数。如果量表间的相关系数随个体不同而异,这表明存在个体—量表互作,这样我们就可见侦测这种互作效应是否存在。然后(如果不存在个体—量表互作的话),我们再求得这些矩阵在个体上的均值,但这个方法和第二个方法具有相同的局限,即只有相对较多的概念才能得到稳健的解。

(5)“外延法”。在这个方法中,我们三个模态中的任何两个进行组合,分别形成一个个案,然后再把它们“外延”成单个数据矩阵。最常见的外延法是,把每个个体—概念答案看做是一个独立个案。然后利用个体—概念(外延模态)×量表的数据矩阵,计算量表间的相关系数。马圭尔的结论是,这是计算量表间相关系数的最佳、最方便的方法。不过,库宾里克和法尔(Kubiniec & Farr,1971:533)曾指出,因为“我们只研究了被试—概念个案上的共变,因而就不可能考察概念间的结构差异”,我们忽略了概念—量表互作。而且,迈尔伯格和比恩(Mayerberg & Bean, 1978)采用两种不

同的外延方法(一种方法如上所述,另一种方法是把每一概念—量表组合作为一个个案),并证明,我们可以从相同的数据中求得不同类型的因子。

对各种 SD 数据的简化方法的上述综述,清楚地表明,随着我们使用的方法不同,我们就会排除特定的变异源。而且,方法不同,分析单位(如被试、概念)也可能会不同。毫不奇怪的是,不同方法可能而且的确生成显著不同的相关系数,甚至到了改变符号的地步。(Bynner & Coxhead, 1979)这样,在 SD 中所使用的量表的意义结构,有关它们的结论在很大程度上取决于所采用的方法。马圭尔(Maguire, 1973),伯恩纳和考克斯黑德(Bynner & Coxhead, 1979),考克斯黑德和伯恩纳(Coxhead & Bynner, 1981),迈尔伯格和比恩(Mayerberg & Bean, 1978)的示例,为我们提供了有关这种状况的、令人惊讶的证据。

### 总结性评述

上述相当有限的综述应足以提醒大家,真实、有意义的 SD 应用伴随着非常复杂的情形。这是我们的希望。这样的情形也有力地表明,任何技术的有意义应用,远不止于了解它们的建构和实施原理。在原理上,比 SD 简单的技术并不多。这也可能是我们经常误用它的原因。

其他任务除外,一个基本的任务是决定所用量表的维度,以及最适合回答所研究问题的分析类型;这是复杂性出现的原因。由综述可见,十分明显的是,对评估应用 SD 的报告而言,对因子分析的初步理解是一个基本要求。而且,尽管我们对分析方法的述评是简要的,但它们传达出一个信息:如果我们想在各种简化数据的程序中作出明智的选择,就必须彻底理解 SD 所试图解剖的实际问题、掌握测量原理的相关知识、熟悉相关的分析技术。

## 访 谈

在我们日常生活中,访谈无所不在。研究语境下,在收集信息的过程中,诉诸一问一答是再自然不过的事情。我们通常使用访谈来收集关于"事实"、意见、态度、行为等方面的信息。在访谈的过程中,问答顺序需要遵守很多特殊规则,外行很难看出来它们的

结构和可能的效应;访谈的高度流行一部分也取决于它和有问有答的日常活动的、众所周知的相似性。宾厄姆(Walter Van D. Bingham)和摩尔(Bruce V. Moore)把访谈刻画为一种“带着一个目的的交谈”(Bingham & Moore, 1941:1)。

大多数关于访谈定义的内核,可以以卡恩和坎内尔(Robert L. Kahn & Charles F. Cannell,1957:16)的定义为示例,即访谈是:

> 一个专业化的言语互动模式——因一个特殊目的而起,关注一些特定的内容,并随时排除外来的干扰。而且,访谈是一个访问员与被访者的角色关系高度专业化的互动模式,它的具体特征在一定程度上取决于访谈的目的和特征。

访谈可能服务于各种各样的目的,它的进程或形态也因此会发生相应地变化。举例来说,下列方面是访谈的主要应用领域,它说明一种访谈可能采取不同的形态:人事(例如,遴选、评价);卫生(例如,病史、诊断);新闻(例如,新闻采访、民意调查);法律(例如,证词);营销(例如,购物习惯、产品偏好);社会行为研究(例如,调查态度、志向、性行为)。

我们对访谈和访问的讨论将集中于它在研究中的应用,这里,“研究”一词是在最宽泛意义上使用的(例如,理论检验、民意测验、项目评估、市场营销)。正如本章所涉及的其他方法一样,我们的讲解是介绍性质的,也没有穷尽所有方面。有关访谈和其他专门方法的一般讨论,请参阅布拉德伯恩等(Bradburn et al, 1979),布伦纳(Brenner, 1978,1981b),坎内尔和卡恩(Cannell & Kahn, 1968),坎内尔等人(Cannell et al, 1981),戈登(Gorden, 1975),卡恩和坎内尔(Kahn & Cannell, 1957),以及米勒和坎内尔(Miller & Cannell, 1988)。

访问可以采用面对面或通过电话两种方式进行。近年来,使用电话有了显著的增加,而且毫无疑问,这种增加还将继续。这一方面是由于技术进步,另一方面是因为一个事实:电话访问更经济、更快捷,因而让我们能够以更及时的方式来抓住各种“热点”问题(例如,对投票行为的预期民意调查)。

由于电话沟通常常容易受到各种限制,没有神韵,特别是缺乏各种非语言的线索,因此,在研究复杂、私密、需要深度追问的问题时,我们更倾向于采用面对面访问。而且,一般来说,面对面访谈的应答率也比电话访谈高出几个百分点。因此,我们主要集中讨

论面对面访问。有关电话访谈的特殊性、它和面对面访问的比较等方面的信息,参见坎内尔(Cannell, 1985b)、弗雷(Frey, 1989)、格罗乌斯和卡恩(Groves & Kahn,1979)、拉夫拉卡斯(Lavrakas, 1987)、舒曼和凯尔顿(Schuman & Kalton, 1985),以及温伯格(Weinberg, 1983)。

## 访谈与问卷

对比访谈与问卷(一种替代访谈的最常见的选项)是一项值得做的工作。所谓"问卷"是指采用纸笔收集信息的工具,常常是被访者自填的。

让我们首先看看问卷对访谈的优势。众所周知,一般来说,问卷的成本低,耗时少,在遴选、培训和监管工作人员等方面的要求也较松。而且,与访谈相比,可邮递问卷一般对研究总体具有更宽的覆盖面。有时候,邮递问卷是接触偏远地方或特殊总体的被访者的唯一手段。

和访谈相比,问卷更统一和标准化,受各种偏差的影响也较小。这些偏差可能源自偏离指示,也可能源自实施方法(在有些类型的访谈中比较常见),更不用提和访问员效应相联系的潜在偏差(参见下文)。最后,使用问卷,保密和匿名能更有效地得到保证。

让我们来看看访谈比问卷具有的优势。首先应注意的是,有些研究领域、有些类型的信息,使用访谈更自然一些。书面问题、陈述和应答都是有限的,而且具有限定效应,访谈则可以探讨更复杂的问题,也可以进行更全面的探讨,为追问参与者的应答提供了机会,在探讨全新的问题上它也更具有灵活性和主动性。在有些情形下,访谈是获得所求信息的唯一可行方式(例如,关于儿童或文盲的信息)。

其次,发生在访问员与被访者之间的互动,让我们拥有可以更多的机会来激励被访者提供更准确的答案、处置各种误差源(例如,误解导语、题器措辞、术语定义等);使用问卷的时候,我们常常无法检测它们。同时,访谈情境也会让被访者难以拒绝回答某些提问或拒访。

而且,在展示题器、题器顺序、排除不相干的题器等方面,访问员可以拥有更多的控制。最后,在访谈情境中,对被访者的观察能够提供有价值的信息和洞见,一般来说,它们是使用问卷所无法获取的。

总而言之，每一种方法都有自己的优势和劣势，我们需要根据所研究的具体现象、特定目的、既定情境、特定来源、被访者等方面，选择它们当中较合适的一个。

## 访谈结构

与其他类型的面对面沟通不同，访谈的启动、执行和中止，受一些规则的制约；在大多数情形下，访问员直白或隐含地制订这些规则。研究性访谈尤其如此，它是陌生人之间的交谈，为的是访问员所设定的一个特殊目的。访问员不仅要保证被访者的合作，而且要激励他们竭尽所能地、诚实地回答向他们提出的问题。而且，我们还要求访问员控制访谈，如果情境需要的话，进行追问、转移和再转移话题、鼓励“恰当”的应答、阻拦“不相干”的应答。

不过，在结构上，不同访谈之间可能会有很大的不同。提问既可以是毫无结构的、松散的、没有导语的，也可以高度指导性的、封闭的。同理，应答也可以是开放式的、没有格式的，也可以是高度结构化的、强迫选择的。访谈的结构性越强，产生清晰、专注的沟通的可能性就越大。我们的讲解主要关注的是相对结构化的访谈，也称做“标准化”或“指导性”访谈。①

## 测量误差与应答效应

与其他测量工具一样，评估访谈的信度和效度的基础是识别各种不同的系统和非系统误差源（详细讨论参见第 5 章）。非系统误差除外，我们可以把访谈中的潜在系统误差分为三个主要类别：访问员（例如，背景变量、期望、追问误差）；任务（例如，措辞、格式）；被访者（例如，背景、态度、应答风格）。当然，在对应答的影响上，这三类因素有可能存在互作效应。

我们将在下一节讨论访问员效应。在一定程度上，很多被访者效应、任务效应和数据采集的其他方式（例如，问卷）相干，因此，在下一节，我们将更全面地讨论它们的潜在效应。

## 访问员效应

访问员效应是不可避免的，正如卡恩和坎内尔（Kahn &

① 非结构访谈的例子，参见洛夫兰德（Lofland, 1971）和米什勒（Mishler, 1986）。

Cannell, 1957：195）所指出的：

> 我们只能把访谈看做是一个互动过程，除此之外，别无他法。字面上来说，互动是指一个人正在影响他人，并以各种不同方式对他人作出反应。因此，说我们想要一个没有访问员效应的访谈，这句话本身就是一个矛盾。

因此，问题不是"是否存在访问员效应"，而是"什么是访问员效应"。而且，我们是否可以识别不同类型的访问员效应？是否存在可以让一些（或全部）访问员效应最小化的方法？[①]

一般来说，我们可以区分"角色有关的效应"与"角色无关的效应"。如名称所示，前者是指因访问员角色的特定方面以及扮演角色的方式所带来的效应。另一方面，角色无关的效应是指主要因访问员的属性（最宽泛的意义上）所带来的效应。我们将分别讨论这两个效应。

## 角色有关的效应

看起来恰恰相反，一般来说，访问员的角色得到很好的规定和界定（有关访问员和被访者角色的详细讨论，参见 Brenner 1978, 1981a, 1982）。大多数时候，我们要求访问员遵守访谈的标准化结构[②]，他（或她）的主要角色是创造各种最佳条件，刺激和激励被访者提供相干的、准确的答案。

因此，坎内尔等人（Cannell et al, 1981：389）指出："访问员具有操纵或扭曲答案的潜能，这引发了各种方法的产生，旨在控制访问员对答案的影响。"这些方法包括行为规范、反馈类型、反馈的适用条件、追问策略、导语的恰当使用。（有关文献，参见 Brenner, 1982; Cannell, 1985a; Cannell et al, 1981; Fowler & Mangione, 1990; O'Muircheartaigh, 1977; Sudman & Bradburn, 1974）

因访问员的角色行为而起的应答效应，是指访问员的角色需求与介入的实际行为之间的差距。在一定的程度上，它和访问员的能力，以及（/或）扮演规定角色的倾向有关（Brenner, 1981a;

---

① 我们建议大家在阅读本章时同时阅读第 11 章的相关部分和访谈者效应相关的章节，请参阅第 11 章"研究者"一节。

② 当然，这不是一种"全或无"的情境，而随着访谈的结构化程度不同，它也会发生变化。在极端的情形下，一个高度结构化的访谈可能类似于一个刺激—反应情境。（Brenner, 1982）

Sudman & Bradburn, 1974)。其他问题除外,研究发现:(a)访问员经常改变题器,并不按照书面题器提问(参见 Bradburn et al,1979;Brenner, 1982; Martin, 1983; Schuman & Kalton, 1985);(b)访问员可能会对答案作出不同的追问和反应(Cannell & Kahn, 1968;Martin, 1983);(c)访问员经常无效地给出反馈,特别是不加区分地给予正反馈(Martin, 1983)。其他偏差效应的示例,参见:Kahn & Cannell, 1957;Hyman. et al, 1954。

综合正反两方面,萨德曼和布拉德伯恩(Sudman & Bradburn, 1974)认为,访问员和应答效应没有任务效应重要。但布伦纳(Brenner,1982:135)在综述了有关角色受限的访问员特征的证据之后,得到的结论是,出现偏差的可能性很大,“在任何具体的调查中,我们都有必要来评估访问员扭曲应答过程的可能程度”。

### 角色无关的效应

角色无关的效应是指源自访问员背景和心理属性的效应。

#### 背景属性

我们经常研究访问员背景属性所带来的效应,包括种族、性别、年龄和社会经济地位(SES)。总的来说,性别和 SES 看起来并不影响应答,但后者的效应有时候是模糊的,特别是因为它倾向于和种族的效应相混杂(Hagenaars & Heinen, 1982)。

另一方面,我们已经发现,当问题与种族和性别特征具有密切联系时(例如,种族态度、性别刻板印象),它们就会影响应答。(有关综述,参见 Cannell & Kahn, 1968; Hagenaars & Heinen, 1982;Schuman & Kalton, 1985; Sudman & Bradburn, 1974)

#### 心理属性

我们把访问员的人格特质、态度、价值观、意见、期望等归类到“心理属性”下面。经常有人主张,访问员的态度、价值观和意见会直接影响应答;或具有间接效应,体现在记录和转录中的错误之上。至少有一部分证据表明,对于特定问题,访问员的意见与被访者的应答之间存在关联(参见 Bingham & Moore, 1941; Cannell & Kahn, 1968; Cantril, 1944; Erdos, 1970)。但是,舒曼和凯尔顿(Schuman & Kalton,1985)认为,几乎没有直接的证据表明,访问员的意识形态会带来偏差。

在讨论访问员遴选问题的文献中,充斥着很多有关受欢迎的人格特质的建议,比如诚实、适应性、和气(参见 Gorden, 1975; Lin, 1976; Sheatsley, 1951; Weingerg, 1983)。但大多数建议的基础是常识,更像是逸闻趣事,几乎没有系统性研究的支持。

在广泛的综述之后,哈根纳斯和海嫩(Hagenaars & Heinen, 1982:126)的结论是:"一般而言,访问员和角色无关的属性并不具有应答效应;只是在特定的情形下,我们才能期待它们出现。"

## 任务效应

下列方面可以归入"任务"类别:题器或刺激(例如,如何呈现、如何措辞);应答(例如,开放—封闭、选项数量);执行程序(例如,反馈、如何处理不太情愿的被访者)。

关于这些主题以及相关话题已经有了很多的讨论(参见第 11 章)。为了避免大幅度的跳跃,我们决定只关注一些有关题器及其应答模态的话题。一个事实蛰伏在我们所讨论的、范围广泛的主题之下:即使我们作了如此严格的限定,我们也只能抓住它的表面现象。对这里所涉及的各种问题的更详细的讨论,参见:Cannell & Kahn, 1968; Cantril, 1944; Converse & Presser, 1986; Gable, 1986; Kahn & Cannell, 1957; Molenaar, 1982; Nunnally, 1978; Schuman & Kalton, 1985; Schuman & Presser, 1981; Sudman & Bradburn, 1974。

### 题器和应答模态

考虑到题器措辞的关键作用,我们将从对这个主题的泛泛讨论开始,随后再讲解题器和应答模态的一些具体方面。

#### 题器措辞

众所周知,一个题器的措辞在很大程度上决定了我们所能得到的答案的种类。下面一个示例来自萨德曼和布拉德伯恩(Sudman & Bradburn, 1982:1)所编撰的一个故事。

> 曾经有两个牧师,一个来自多明我会,一个来自耶稣会,正在讨论吸烟和祷告同时进行是不是一种罪孽。由于没有形成一个结论,他们回去询问各自的上司。第二周他们又见面了。多明我会的牧师问:"你的上司怎么说的?"耶稣会的牧师

答:“他说这没有罪。”多明我会的牧师回应道:“太有趣了,我的上司说这有罪。”耶稣会的牧师问:“你问他什么?”答曰:“我问他,祷告时吸烟可以吗?”耶稣会的牧师说道:“哦,我问上司,吸烟时祷告可以吗?”

讨论如何编写题器、良好的题器措辞具有哪些要素的文献已经汗牛充栋(参见 Cantril, 1944; Converse & Presser, 1986; Converse & Schuman, 1984; Hogarth, 1982; Miller, 1983; Oppenheim, 1966; Payne, 1951; Sheatsley, 1983; Sudman & Bradburn, 1982; Turner & Martin, 1984: 第一卷,第九章)。正如佩恩(Payne,1951)的经典著作《提问的艺术》的书名所传达的那样,题器编写或题器措辞具有相当的艺术成分。如果没有创造性的要素,没有跳跃思维的能力,没有简明扼要的沟通,我们很难想象能够出现遣词优美的题器或选项。

影响应答的主导因素是,对被访者而言,词汇、短语及其组合是含义有别的。对研究者来说,有些词汇或题器的意义是“非常清晰的”,但被访者可能不理解,或者理解成不同的意义。

这种情形的一个有趣示例,来自舒曼和凯尔顿(Schuman & Kalton, 1985,引自舒曼的一个研究)。在一个访谈的过程中,很多题器呈现给被访者,要求他们表明自己是赞同还是不赞同,其中包括:“尽管人们有各种说法,但很多普通人过得越来越差,而不是越来越好。”(Schuman & Kalton, 1985: 642)在这之前,很多人曾经多次使用过这个题器[1],因此,他们假设它应当不成问题,没有经过前测,他们就把它放在访谈提纲的终稿之中。”(Schuman & Kalton, 1985: 642)结果是,访问员指出:

> 在整个持续一个多小时的访谈中,“很多普通人”这个问题肯定是问题最多的一个,主要因为许多美国人并不熟悉 lot 一词的用法。对这个题器存在各种各样的理解,如“很多普通男人”“宅地的面积”,甚至有一个人把它理解成“墓地”!(Schuman & Kalton, 1985: 642-643)

前一个例子关注的是同一个题器的意义或阐释。毋庸赘言,当题器使用不同的措辞时,一般来说,情形会变得更复杂。因为措辞效应的作用方式是细微难料的,“认为一个‘相同’的题器可以具

---

① 顺便提一下,这个题器来自斯罗尔(Srole, 1956)流行的“失范量表”。

有别样的措辞,这种认识是错误的。措辞上的任何变化都将改变题器的意义”(DeLamater, 1982:23)。甚至当我们使用相同的选项,但改变选项的顺序时,意义都有可能受到影响。下面我们将针对每一种情形举一个例子,它们均来自舒曼和普雷瑟(Schuman & Presser, 1981)的一部优秀著作,在这本书中,大家可以找到大量信息,讨论有关题器的格式、措辞和语境的各种实验。

“同一”问题的两种形式(Schuman & Presser,1981:281):

| “禁止”形式 | “允许”形式 |
| --- | --- |
| 您认为美国应该禁止宣扬共产主义的公共演讲吗? | 您认为美国应该允许宣扬共产主义的公共演讲吗? |

选项的次序(Schuman & Presser,1981:60):

| “石油充裕”说在前 | “石油充裕”说在后 |
| --- | --- |
| 有人说,我们还有大量石油,足以让我们再用25年。也有人说,按照我们现在的用油速度,只要15年左右我们就会把石油用光。您认为哪种观点更正确? | 有人说,按照我们现在的用油速度,只要15年左右我们就会把石油用光。也有人说,我们还有大量石油,足以让我们再用25年。您认为哪种观点更正确? |

撇开细节不谈,需要指明的是,前两个例子当中的每一个都会得到不同的应答模式(即赞同/不赞同的百分比)。

在探讨题器措辞的方向性时,赖泽等人(Reiser et al,1986)指出,面对表达相同观点的陈述时,人们更倾向于赞同负面措辞的陈述,而不是不赞同正面措辞的陈述(即赞同“大多数人不可信”,而不是不赞同“大多数人可信”)。

信息呈现的格式也会影响偏好和决策,麦克尼尔等人(McNeil et al,1982)的研究就是一个示例。研究人员让患者、医生和研究生(具有统计学和决策论的坚实背景)想象自己患了肺癌,然后依据提供给他们的信息,让他们在两种疗法中进行选择。这些信息的一个方面是基于死亡率或存活率进行描述的(例如,10%的死亡率或90%的存活率)。无论背景如何,很多被访者都偏好“以生存概率而不是以死亡概率”描述疗效的治疗方案(McNeil et al, 1982:1259)。

## 开放—封闭之分

题器结构的一个主要方面是指我们想要的应答模态。它的一端是开放题(也称"自由形式"或"无结构"),要求开放的回答;另一端是封闭题(也称"必选"),要求被访者在所提供的一组答案中进行选择。有各种不同的格式可以引发必选应答,包括是/否、多项选择、表单、评分量表以及各式各样的赞同—不赞同格式。当然,题器结构可以在这两端之间进行变动。

开放题还是封闭题?这是一个备受争议的主题。部分原因是,有些学者和研究者倾向于把这两种题型和研究的不同取向联系起来。他们认为,开放题适合于定性研究,封闭题更适合定量研究。一个格式是否比另一种格式更适合,在何种情形下更适合,对它们的研究尝试,参阅舒曼和普雷瑟(Schuman & Presser, 1981),以及康弗斯和普雷瑟(Converse & Presser, 1986)。

## 选项数量

探讨这个问题的大多数研究都集中在评分量表上,尽管有些结论也可能适合于其他题型。在评分量表中所采用的备择选项的数量,变化幅度很大,既有建议 2 个或 3 个(参见 Jacoby & Matell, 1971),也有建议在有些条件下为 25 个(参见 Guilford, 1954)。碰到 100 个潜在备择选项的量表(例如,以百分数为基础的量表)或者对备择选项没有数量限制的量表,甚至也不算什么不同寻常。

许多研究者都考察了量表点的数量对信度、效度、复原性和被访者偏好的效应(参见 Comrey & Montag, 1982; Garner, 1960; Green & Rao, 1970; Komorita & Graham, 1965; Lissitz & Green, 1975; Matell & Jacoby, 1971, 1972; McKelvie, 1978; Ramsay, 1973),但结论并不一致,甚至还有一些争议;看起来,最常见的建议是:一个量表应由 5 到 9 个点构成(参见 Cox, 1980; Gable, 1986; Molenaar, 1982; Nunnally, 1978)。

## 中间类别

是否应当包含一个中间类别,作为备择选项中的一个?对这个问题的回答莫衷一是。首先,对不同的人、不同的题器类型、不同的研究领域而言,中间类别的含义完全不同。下面是中间类别

所采用的术语或类似术语:不知道、中立、犹豫不决、没有意见、没有差异、没有评论、不承诺、中立位置。很明显,它们并不是同义词。下面我们将简要小结一下有关使用这些术语的一些发现。

### 逻辑上的中立位置

对有些题型而言,设一个中立位置显得合乎逻辑。诸如"适量""和现在差不多""路中间""不轻也不重"等答案,很显然就是一些题器的逻辑上的中立位置。有人发现(参见 Gable, 1986; Molenaar, 1982; Schuman & Presser, 1981),当有一个中立位置作为备择选项时,选择它的被访者的百分比,会高于没有这个中立位置,而主动给出这个选项的被访者比例。

我们必须认识到,当我们提供一个中立位置时,人们选择它的理由是各种各样的。有些人选择它,可能是因为它是一个容易或快捷的出路;有些人选择它,可能是因为它是应对焦虑的一种手段,这种焦虑来自要求他们回答一个提问,但他们碰巧几乎或完全不了解,或者他们完全没有考虑过;有些人选择它,可能是因为他们根本没有理解提问。这种状况给分析和阐释研究结果带来了困难(对中立位置的详尽的讨论,参见 Schuman & Presser, 1981:第 6 章)。

### 不知道或没意见

当我们试图处置"无态度"(Converse, 1970)或"假意见"(Bishop et al, 1980)时,我们就会包括一个"不知道(DK)"或"没意见(NO)"的选项。所谓"无态度"或"假意见"是指,当人们面对完全不熟悉的评估对象和话题时,他们发表意见或表达态度的意向性。它的一个生动展示是人们对根本不存在的评估对象或话题而发表意见或表达态度的一种倾向。例如,有研究表明,有相当比例的被访者会对虚假的(a)种族群体、(b)国会法案、(c)宪法修正案表达自己的态度(最近一项研究,参见 Bishop et al, 1986。有关无态度的研究和争论的综述和评估,参见 Smith, 1984)。

研究表明,当 DK 不是一个可能选项时,大约 10% 的被访者会选择说"不知道"。但是,当 DK 是一个相同题器或提问的一个可能选项时,大约 30% 的被访者倾向于选择它。类似地,一个评分量表的中间类别(例如,7 点量表中的 4)也会吸引较多的被访者(Aldrich et al, 1982)。

研究表明，选择 DK 与很多被访者变量（如受教育水平，参见 Converse, 1976）、题器歧义（参见 Coombs & Coombs, 1976）和特定主题有关联（对这些问题和更宽泛的题器措辞论题的详尽探讨，参见 Schuman & Presser, 1981）。

必须记住，对于不同的被访者和不同的研究者而言，DK 具有不同的意义。赋予 DK 的意义包括：不了解、忽视、无差别、歧义。大多数研究者把 DK 和 NO 当作同义词使用，有些研究者则严格进行区分，考虑到这个事实，事情就变得更复杂了。

毫不奇怪，很多研究者（参见 Andrich, 1978; Bock & Jones, 1968; Converse & Presser, 1986）建议，一般来说，尽量不用中间类别或慎重使用它。

### 题器顺序

另一个引起争议的论题是题器顺序的潜在效应，也称“顺序效应”。一方面，简单改变一下一个题器在问卷或访谈提纲上的总位置，似乎并不产生任何大一点的效应（Molenaar, 1982）。另一方面，当我们考虑到题器所处的语境时，诸如重点、连贯和对比等顺序效应变得十分明显（Cantril, 1944; Kahn & Cannell, 1957; Molenaar, 1982; Schuman & Presser, 1981）。

正如舒曼和普雷瑟（Schuman & Presser, 1981）所指出的，当所包含的题器是针对一组紧密关联的话题时，顺序效应（更具体地说，是情境顺序效应）就更有可能发生。一个问卷中的题器常常是按照话题编排的，这个事实让问题变得更复杂。而且，一般性题器似乎比具体性题器对顺序效应更敏感。有些调查研究者的行事方式好像不存在题器顺序效应或效应很小似的，为了引起他们的注意，舒曼和普雷瑟（Schuman & Presser, 1981:74）曾警告说：“考虑到过去的研究和我们自己的实验，我们必须挑战这些便于行事的假定。”

## 被访者效应

如前所述，我们可以在“被试效应”的更宽泛语境下来讨论被访者效应。前者我们将在第 11 章中作深入探讨，这里我们仅对被访者效应做一点一般性的观察，特别是它们在应答题器（无论是在

访谈中,还是在问卷中)时的情形。

首先,我们请大家注意前面所作的评述,即我们并不能清晰地把访谈过程中的有些方面归类到三个类别(即访问员、任务和被访者)之一,而这三者是我们现在所关注的。而且,在它们对应答的效应上,不同类别之间还可能存在互作。这里,我们并不想详细讲解互作的含义①,就当前的目的而言,我们仅仅指明,它是指两个或多个变量的联合效应。互作的一个例子是,当面对复杂的题器时,超常比例的小学水平的被访者选择"不知道"这个选项,大学水平的被访者就不是这样。

在被访者角色扮演的各个方面中,最重要和最普遍的方面可能就是"自我呈现",它是指被访者一方希望在特殊的光照下向研究者展示自己,以造成一个特别的印象。虽然"在其他条件保持不变的前提下,一般假定,人们会以降低人际不适或社会不适的方式行事,或者尽可能给别人留下好印象"(Sudman & Bradburn, 1974:9),但自我呈现并不必总是正面的(有关示例,参见第11章关于叛逆被试的讨论)。

由此可推论,情境的特性会影响人们对自我呈现的关注。因此,在面对面访谈的情境下,被访者可能会关注他们所投射的形象,而在匿名填答一份邮递问卷时则不会。不过,相应地,这也取决于任务因素。例如,面对面访谈的应答与匿名自填问卷的应答之间的差异,将取决于题器的性质。在这两种情境下,比较平淡的题器很可能引发相似的应答;"令人为难的"题器则可能造成不同的应答。在面对面的访谈中,令人为难的题器更可能带来社会赞许的应答,但具体效应还取决于访问员的地位(例如,教师、精神科医生、大学助教)。

前面所论及的问题似乎还不够复杂;被访者可以承担多种角色,这些角色甚至可能相互冲突,这个事实才让问题变得更复杂。这种情形如何会出现?它带来什么后果?对这些问题的讨论参见第11章。

在前一节我们已经讨论过,被访者对题器的理解将影响应答的性质(也可参见第11章,理解假象)。同时,被访者是否可以接近我们想要收集的信息,也会影响应答(Cannell et al, 1981;Cannell &

---

① 直观的介绍参见第10章,详细的讨论参见第20、21章。

Kahn, 1968; Converse & Presser, 1986; DeLamater, 1982; Kahn & Cannell, 1957; Sudman & Bradburn, 1974)。当然,由于各种各样的原因,包括健忘、压抑、不能或不愿言表,信息的可接近度是不同的,有些信息甚至完全无法接近。它的潜在后果包括低估(假阴性)、高报(假阳性)和应答的扭曲。

### 应答风格

应答风格(也称做"应答定势")是无视题器内容而给出答案的倾向。研究最多的两种应答风格是随声附和和社会赞许性。

随声附和。"随声附和"或"唯唯诺诺"最早由克伦巴赫提出,它是指被访者答"是"多于答"否"的倾向(Cronbach, 1946)。库奇和凯尼斯顿(Couch & Keniston,1960:169)认为:"这带来的一个后果是,问卷题器似乎被一种倾向所'霸占',无论它们的内容是什么,它们彼此之间都正相关。"第4章,在讨论权威主义的一个度量(F量表)时,我们曾经给出了有关这种应答风格的研究的参考文献。

一般来说,随声附和可作为混杂误差的一个来源,特别是当题器或刺激存在歧义时(参见 Brenner, 1981a; Cannell et al., 1981; Converse & Presser, 1986; DeLamater, 1982; Jackson, 1967; Messick, 1967; Schuman & Presser, 1981; Wiggins, 1973)。[①] 另外,有些初步证据表明,受教育程度较低、对任务投入较少的被访者,随声附和的现象发生的频率更高一点(Schuman & Presser, 1981)。

社会赞许性。爱德华首创"社会赞许性"这种应答风格,它是指被访者将自己较好的一面呈现给研究者或访问员的倾向(Edwards, 1957a),因此,也称做"自我赞许性"(Nunnally, 1978)。虽然社会赞许性的效应规模及其一致性还是一个备受争论的主题,但有研究发现,社会赞许性会影响对问卷和访谈的应答(参见 Brenner, 1981a; Cannell & Kahn, 1968; Cannell et al, 1981; DeLamater, 1982; Edwards, 1967a, 1967b; Sudman & Bradburn, 1974; Wiggins, 1973)。

---

① 但是,纽纳利(Nunnally, 1978)认为:"无论是作为人格的一种度量,还是作为人格和情绪度量中系统性误差的一个来源,附和倾向都无足轻重。"

# 观 察

众所周知,通过观察,我们了解自己所处的物理环境和社会环境。但无论是日常习惯的、看似漫不经心的观察,还是科学研究精心的、系统性的观测,重要的是要认识到,观察是一个主动的过程。这暗含一个特定的参照系,它带来对观察对象("什么")和过程("怎样")的选择。为了说明这一点,波普尔(Popper,1972:259)让读者们参与一个实验,即告诉他们"此时此地,观察"。然后他说道:

> 我希望你们大家互相合作,进行观察!但是,我担心你们当中至少有一些人,不是观察,而是急匆匆地问:"您要我们观察什么?"如果这是您的反应,那么我的实验就成功了。我想举例说明的是,为了能够观察,我们必须在心中有一个明确的问题,一个或许可以通过观察就可以决定的问题。达尔文很清楚这一点,他写道:"所有观察都必须支持或反对某些观点,看不到这一点,是多么奇怪的事情。"(Popper,1972:259)

子曰:饱食终日,无所用心(转引自 Armstrong, 1985:xi)。孔子的这句话,可以作为结语[①]。

有些人认为,观察是"一切科学的基础"(Brandt, 1972:22),但有些人仅把它看做是某些研究策略的基石(参见 Bogdan & Taylor, 1975; Erickson, 1986)。实际上,我们经常把"观察研究"这个术语和"自然研究""田野研究""民族志""人类学研究""非实验研究"和"准实验研究"等术语混用。[②] 这已经造成各种错误的观念,主要的一个谬误是区分观察研究和实验研究(参见 Weick, 1968)。

我们认为,"观察研究"这个术语是个不当的命名,因为它错误地把一种收集数据的方法(即观察)和研究方法或研究设计并列起来,包括实验(参见第 12 章)、准实验(第 13 章)和非实验(参见 14 章)。总之,观察只不过是众多收集数据和测量程序中的一种,正因为如此,它"与使用尺子、测量电流或称重化学品等科学活动属于一个类别"(Fassnacht, 1982:39)。

---

① (阿姆斯特朗没有给出引文出处,译者浏览《论语》全文,可能这是较接近的一句)。

② 参见第 13、14 章。

在社会行为研究中，作为典型的用法，观察关注的是口语或语言行为，这很可能是因为语言更容易编码，而其他行为则较难，包括语言之外的元素（例如，音调、音长）、面部表情、身体动作和人际空间。虽然我们常常把观察应用于“自然生境”[①]，但观察自身内部并不存在将自己局限于这类场景的因素。似乎并没有必要重复说，我们也可以在实验室、在实验条件下、在模拟情境中等进行观察。但考虑到前面所提及的有关观察研究的错误观念，这样说一次还是有必要的。

不出所料，观察程序也有自己的问题。首先，它们一般都比较昂贵。其次，它们对“观察者效应”（例如，期望、晕轮、记录错误）和被观察者一方的反应比较敏感。最后，被观察者的隐私也常常会受到连累（论证的正反两方面的小结，参见 Fassnacht，1982）。

为了理顺观察的目的、情境和目标的多样性，人们提出了一些分类系统（参见 Brandt，1972；Evertson & Green，1986；Medley，1982；Weick，1985；Wiggins，1973）。其中，法斯纳赫特（Fassnacht，1982）提出的分类可能是最细致的。由赫伯特和阿特里奇（Herbert & Attridge，1975）提出来的标准（例如，目的、内容、定义、信度、效度和可操作性）对分类系统的使用者和开发者，或许都会有所帮助。

与前面章节所讨论的其他方法一样，我们对观察法的讲解必然是挂一漏万的。事实上，我们只讲解了一些有关系统性观察的问题。[②] 为了说明系统性观察是什么，如何区分它与其他形式的观察，我们采用韦克（Weick，1985：568）的定义，或许会有助于大家的理解：

> 系统性观察是指持续、详尽、有条不紊地观察和转述各种社会情境，这些社会情境和自己自然发生的背景有联系。

我们需要指明这个定义的几个要点。走马观花、无意识、无计划和无组织的观察，都被排除在考量之外。所谓“转述”是指观察者选择性地、主动地阐释所观察到的事物。“社会情境”暗含观察的对象包含三个要素——行动者、场景和所介入的活动。如果大家想深入了解，请阅读韦克（Weick，1968）对这个定义所含的七个

---

① 事实上，“现场”也是对“观察”对定义的一部分（参见 Weick，1985：968）。

② 对照系统性观察，对其他形式的观察（例如，生态学观察、民族志观察）的讲解，参见埃弗森和格林（Evertson & Green，1986）和梅德利（Medley，1982）。

要素的详细讲解。在上述讨论的基础上,下面我们转向对观察数据的讨论。

## 观察数据[①]

我们能够采集的数据类型,记录数据所采用的工具是多种多样的,就如同观察中包含不同的行动者、场景、对象、内容和实质一样。勃兰特(Brandt,1972)曾对这个主题进行过最全面的讲解,他提出一个分类,包括下列类别:

> 1. 叙述:它包含"只复制行为事件,保持它们原初发生时的方式和顺序"的数据(同上:80)。例如,轶事、样本记录、田野笔记、生态描述、信件、日记。
>
> 2. 评分:例如,数字、图式。
>
> 3. 清单:"把范围限定到行为和情境的特定方面,观察者对它们也容易形成共识"(同上:81),包括的事物有:静态描述(例如,年龄、性别)、行为清单(例如,互动分析、类别系统、符号系统)、活动日志、离散事件记录、特质指标清单。

我们的讲解局限于系统性观察的一些方面,因此,我们只讨论评分和清单。

## 评　分

在社会行为研究中,评分量表十分流行(参见本章的开篇部分);顺理成章的是,它们也大量应用于观察提纲之中(Brandt, 1972; Fassnacht, 1982)。在观察提纲中,评分量表的典型应用是:对行为评分(例如,它们的强度、频率)、依据不同标准对互动进行评估、评估环境或场景(例如,工作场所、操场、教室)。

我们在前面的章节已经讨论过使用评分量表中的一些问题,这里不再重复,但提醒大家注意,评分者在其中起着关键作用(例如,背景特征、参照系、偏见、晕轮效应)。在平常的应用中,我们是在观察快结束的时候,而不是在观察进程中进行评分,这样会加重和应用评分量表相联系的各种问题。而且,在少量行为样本的基

---

① 里默尔(Leamer, 1988: 493)曾说过:"我确实不喜欢把观察数据组成一个词组,这不等于是观察观察吗?"我们同意他的观点,但由于这个术语广为流行,我们还必须使用它。

础上，频繁使用总评分，也是值得商榷的（Foster & Cone，1986）。

## 清　单

为了确保能够在一个特定情境中注意和记录特定的事物，我们会使用清单，这种做法已经具有上百年的历史（Brandt，1972：94）。作为一种观察程序，我们使用它们来记录事先规定的行为和互动的发生和频次（有时记录）。清单具有很多种类和格式，最常用的清单类型，要么是由符号系统构成，要么是由类别系统构成（Medley & Mitzel，1963）。

### 符号系统

一个符号系统，也称做“特征系统”（参见 Fassnacht，1982），包含一个我们感兴趣的行为、事故或事件的列表。在一个既定的观察期内，只要符合列表中的一项，观察者就需要记录行为的发生，有时还包括它们的频次。其背后的假定是，观察到的行为或“事件被看做是研究者感兴趣的一些特征是否存在的指标”（Medley，1982：1842）。这样，“符号的作用就和客观测验中的‘及格/不及格’题器一样”（同上）。

观察者的任务是记录对应于清单所列的行为；清单上没有的行为，就不记录。符号系统可以包含一组较广范围的行为，实际上，它可以同时容纳几个行为域。这样的话，观察者就需要随时识别并注意到全部行为。不过，一份横跨很多领域的行为清单，并不必然会给观察者带来过重的负担，因为一般来说，符号的定义足够狭窄和明晰，可以让观察者快速、简便地进行识别。而且，相对不常出现的行为，也常常包含在清单上。因此，往往会出现一个时段，观察到的行为都是不相干的。此时，有些观察者会有困扰，他们更容易分神，出现“观察者漂移”的现象①（Foster & Cone，1986；Medley & Mitzel，1963）。

### 类别系统

一个类别系统是由一组互斥、穷尽的类别所构成，我们用它来

① 观察者漂移“是指，随着时间的推移，观察者倾向于改变他们应用行为定义的方式”（Kazdin，1977：143）。

对一个领域中所观察到的每一个行为进行分类。观察者的基本任务是把相干行为分配到提纲所提供的类别中去。我们假定:在观察期内,这些行为的每一次出现都会被观察到,而且归类到唯一的类别之中。

类别的数量必须要少(例如,少于 10 个,参见 Medley & Mitzel, 1963);它们应当范围有限、足够明晰,以便观察者能够比较快捷、容易地识别相干行为并归类。无需赘言,一定会发生选择,但请注意,类别已经提供给了观察者,这样,我们就局限在可以归类到这些类别当中的行为。

和符号系统相比,一般来说,类别系统更高级一些,具有更好的理论基础,但也更难以建构。类别的穷尽假定和明晰要求,必然要求该类别系统的作者对研究领域具有一个十分清晰的理解。

## 信　度

与任何度量一样,我们也需要评估观察度量的信度。我们在第 5 章讲解了信度的综述,对古典测量理论给予了强调。观察数据的信度是很多讨论和争论的主题(参见 Berk, 1979; Fleiss, 1986: 第 1 章; Frick & Semmel, 1978; McGaw et al, 1972; Medley & Mitzel, 1963; Mitchell, 1979; Robinson, 1957; Rowley, 1976; Sackett, 1978; Tinsley & Weiss, 1975; Towstopiat, 1984; Weick, 1968)。我们并不想对这些争论的细节进行综述,我们的目的仅仅是介绍一下其中可能涉及的难点和问题。更多的信息,参见上面所列的参考文献。

评估观察数据的信度的最常用方法,是计算"观察者间相合度"(也称"评分者间相合度")[1]。大家都明白,观察者间相合度可以涉及一个特定的误差源,它本身就可能是十分重要的,但它根本不是信度的一个指数,这是一个问题。虽然不断提醒研究者关注这个问题(参见上面所列参考文献)[2],观察者间相合度仍然是最经常取得估值的指数。在很多情形下,它甚至是唯一报告的指数。

---

① 有关准则相合度、观察者内相合度、观察者间相合度的区分,参见 Frick & Semmel, 1978。

② 参见法斯纳赫特(Fassnacht, 1982),他认为古典测量理论的信度应用于观察数据是不恰当的,观察者间一致就足够了。

一般来说,观察者间相合度是指两个(或多个)观察者在编码、评分、分类等方面的一致程度。人们提出过很多“观察者间相合指数”,基本上,他们都以估计观察者之间相合的百分比为目标。在其他条件保持不变的前提下,特定指数的不同之处在于(a)对相合程度的敏感性、(b)是否校正“偶然相合”。

如前所述,在古典测量理论中,我们把信度定义为真值方差与观察值方差之比(参见第5章)。由于通过观察所采集数据的数量和类型,信度估值比较复杂。一方面是因为在任何一段观察期内,一个行为常常被多次抽中采样;一方面是因为在任一时点上,我们可以使用多个观察者;一方面是因为我们可以安排多个观察期。因此,观察中的测量误差可能来自很多源头。例如,所研究的行为、行为样本可能不恰当,行为本身可能发生随机变化,被观察者可能会随着场景的变化而改变,观察的环境发生变化等,都会造成观察者们达不成一致。

观察者间相合度指数只涉及观察者之间的潜在误差。而且,这些误差反映的是观察者在使用观察工具上、在打分上的差异,而不是行为本身的差异。观察者间相合度肯定十分重要,我们应当加以评估,但它并没有涉及更广泛的关切。已经证明,观察者间相合度可能很高,但信度却很低。在这些条件下,低信度的潜在因素包括:总分相同,但单个题器之间的存在不一致;随着场合不同、行为发生变异;就所研究的现象而言,被观察的群体相对同质;观察者漂移。

人们已经提出几种方法(相关系数法和依赖方差分析程序的方法),来评估观察者间的信度(参见 Berk, 1979; Tinsley & Weiss, 1975)。它们有一个相同的缺陷:它们只得出单一指数,因而忽略了观察数据中多种潜在的误差来源(参见上文)。有人认为,能够区分变异的不同来源(例如,场合、观察者、被试、行为)的概括度理论(参见第5章的简短介绍和参考文献),非常适合评估观察者间信度。

## 总结性评述

在本章,我们对社会行为研究中的各种测量方法进行了综述。选择这些方法进行讲解的原因,一是它们的应用流行程度,一是它们可能

暗示有多样性存在。请大家注意,我们的讲解必然有限,既没有穷尽所综述的各种方法,也没有穷尽所提出的问题。

本章是本书第 1 部分的结尾。我们不断提醒大家,请参考在其他部分出现的设计和分析考量,这部分的主题主要还是测量问题。从下章开始,我们转向设计问题。但是,由于研究的各个方面具有内在联系,我们还将回到第一部分所提出的一些问题,在讲解相干的分析技术时,我们也会继续参考第 3 部分。

# 参考文献


Aaker, D. A., & Day, G. S. (1983). *Marketing research*. New York: Wiley.

Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. (1950). *The authoritarian personality*. New York: Harper & Row.

Aldrich, J. H., Niemi, R. G., Rabinowitz, G., & Rohde, D. W. (1982). The measurement of public opinion about public policy: A report on some new issue question formats. *American Journal of Political Science*, *26*, 391-414.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Althauser, R. P. (1974). Inferring validity from the multitrait-multimethod matrix: Another assessment. In H. L. Costner (Ed.), *Sociological methodology 1973-1974* (pp. 106-127). San Francisco: Jossey-Bass.

Althauser, R. P., & Heberlein, T. A. (1970). Validity and the multitrait-multimethod matrix. In E. F. Borgatta & G. W. Bohrnstedt (Eds.), *Sociological methodology 1970* (pp. 151-169). San Francisco: Jossey-Bass.

Alwin, D. F. (1974). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In H. L. Costner (Ed.), *Sociological methodology 1973-1974* (pp. 79-105). San Francisco: Jossey-Bass.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: Author.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: Author.

Anderson, A. B., Basilevsky, A., & Hum, D. P. J. (1983). Measurement: Theory and techniques. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 231-287). New York: Academic Press.

Andrich, D. (1978). A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, *31*, 84-98.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Arbuthnot, J. (1972). Cautionary note on measurement of field independence. *Perceptual and

*Motor Skills*, 35, 479-488.

Armor, D. J. (1972). School and family effects on black and white achievement: A reexamination of the USOE data. In F. Mosteller & D. P. Moynihan (Eds.), *On equality of educational opportunity* (pp. 168-229). New York: Vintage Books.

Armstrong, J. S. (1985). *Long-range forecasting: From crystal ball to computer* (2nd ed.). New York: Wiley.

Arvey, R. D., & Faley, R. H. (1988). *Fairness in selecting employees* (2nd ed.). Reading, MA: Addison-Wesley.

Avison, W. R. (1978). Auxiliary theory and multitrait-multimethod validation: A review of two approaches. *Applied Psychological Measurement*, *2*, 431-449.

Bagozzi, R. P. (1980a). *Causal models in marketing.* New York: Wiley.

Bagozzi, R. P. (1980b). Performance and satisfaction in an industrial sales force: An examination of their antecedents and simultaneity. *Journal of Marketing*, *44*, 65-77.

Bagozzi, R, P., & Fornell, C. (1982). Theoretical concepts, measurement, and meaning. In C. Fornell (Ed.), *A second generation of multivariate analysis* (Vol. 2, pp. 24-38). New York: Praeger.

Barrett, J. P. (1974). The coefficient of determination—some limitations. *The American Statistician*, *28*(1), 19-20.

Bass, B. M. (1955). Authoritarianism or acquiescence? *Journal of Abnormal and Social Psychology*, *51*, 616-623.

Bell, E. T. (1945). *The development of mathematics* (2nd ed.). New York: McGraw-Hill.

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*, 155-162.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency*, 83, 460-472.

Berk, R. A. (Ed.). (1982). *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press.

Berktson, J. (1947). "Cost-Utility" as a measure of the efficiency of a test. *Journal of the American Statistical Association*, *42*, 246-255.

Bingham, W. V. D., & Moore, B. V. (1941). *How to interview.* New York: Harper & Brothers.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-opinions on public affairs. *Public Opinion Quarterly*, *44*, 198-209.

Blalock, H. M. (1971). Causal models involving unmeasured variables in stimulus-response situations. In H. M. Blalock (Ed.), *Causal models in the social sciences* (pp. 335-347). Chicago: Aldine.

Bock, R. D., & Jones, L. V. (1968). *The measurement and prediction of judgment and choice.* San Francisco: Holden-Day.

Bogdan, R., & Taylor, S. J. (1975). *Introduction to qualitative research methods: A phenomenological approach to the social sciences.* New York: Wiley.

Bohrnstedt, G. W. (1983). Measurement. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 69-121). New York: Academic Press.

Bohrnstedt, G. W., & Borgatta, E. F. (1981). Foreward. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 9-20). Newbury Park, CA: Sage.

Borgatta, E. F. (1968). My student, the purist: A lament. *Sociological Quarterly*, *8*, 29-34.

Borgatta, E. F., & Bohrnstedt, G. W. (1981). Levels of measurement. Once over again. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social measurement: Current issues* (pp. 23-37). Newbury Park, CA: Sage.

Bradburn, N. M., Sudman, S., & associates. (1979). *Improving interview method and questionnaire design.* San Francisco: Jossey-Bass.

Brandt, R, M. (1972). *Studying behavior in natural settings.* New York: Holt, Rinehart & Winston.

Brennan, R. L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing.

Brenner, M. (1978). Interviewing: The social phenomenology of a research instrument. In M. Brenner, P. Marsh, & M. Brenner (Eds.), *The social contexts of method* (pp. 122-139). New York: St. Martin's Press.

Brenner, M. (1981a). Patterns of social structure in the research interview. In M. Brenner (Ed.), *Social method and social life* (pp. 115-158). London: Academic Press.

Brenner, M. (Ed.). (1981b). *Social method and social life.* London: Academic Press.

Brenner, M. (1982). Response effects of "role-restricted" characteristics of the interviewer. In W. Dijkstra & J. van der Zouwen (Eds.), *Reaponse behaviour in the survey-interview* (pp. 131-165). London: Academic Press.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*,*54*, 106-148.

Brodbeck, M. (Ed.). (1968). *Readings in the philosophy of the social sciences.* New York: Macmillan.

Brody, J. E. (1983b, May 27). Masters and Johnson defend pioneer sex therapy research. *The New York Times*, p. 13.

Burke, C. J. (1963). Measurement scales and statistical models. In M. H. Marx (Ed.), *Theories in contemporary psychology*, (pp. 147-159). New York: Collier-Macmillan.

Bynner, J., & Coxhead, P. (1979). Some problems in the analysis of semantic differential data. *Human Relations*, *32*, 367-385.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 15, 546-553.

Campbell, D. T.. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Campbell. N. (1952). *What is science*? New York: Dover.

Cannell, C. F. (1985a). Experiments in the improvement of response accuracy. In T. W. Beed & R. J. Stimson (Eds.), *Survey interviewing. Theory and techniques* (pp. 24-62). Sydney:

Allen & Unwin.

Cannell, C. F. (1985b). Interviewing in telephone surveys. In T. W. Beed & R. J. Stimson (Eds.), *Survey interviewing: Theory and techniques* (pp. 63-84). Sydney: Allen & Unwin.

Cannell, C. F., & Kahn, R. L. (1968). Interviewing. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, 2nd ed., pp. 526-595). Reading, MA: Addison-Wesley.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. In S. Leinhardt (Ed.), *Sociological methodology* 1981 (pp. 389-437). San Francisco: Jossey-Bass.

Cantril, H. (1944). *Gauging public opinion.* Princeton: Princeton University Press.

Caplan, R. D., Naidu, R. K., & Tripathi, R. C. (1984). Coping and defense: Constellations vs. components. *Journal of Health and Social Behavior*, *25*, 303-320.

Carroll. L, (1960). *Alice's adventures in wonderland & through the looking-glass.* New York: New American Library.

Carter, L. F. (1971). Inadvertent sociological theory. *Social Forces*, 50, 12-25.

Caws, P. (1959). Definition and measurement in physics. In C. W. Churchman & P. Ratoosh (Eds,), *Measurement: Definitions and theories* (pp. 3-17). New York: Wiley.

Christie, R. & Garcia, J. (1951). Subcultural variation in authoritarian personality. *Journal of Abnormal and Social Psychology*, *46*, 457-469.

Christie, R., Havel, J., & Seidenberg, B. (1958), Is the F scale irreversible? *Journal of Abnormal and Social Psychology*, *56*, 143-159.

Christie, R., & Jahoda, M. (Eds,). (1954), *Studies in the scope and method of "The authoritarian personality.*" Glencoe, IL: Free Press.

Churchman, C. W., & Ratoosh, P. (Eds.). (1959). *Measurement: Definitions and theories.* New York: Wiley.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115-124.

Cliff, N. (1982). What is and isn't measurement. In G. Keren (Ed.), *Statistical and methodological issues in psychology and social sciences research* (pp. 3-38). Hillsdale, NJ: Erlbaum.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, *10*, 637-666.

Cole, N. S. (1981). Bias in testing. *American Psychologist*, *36*, 1067-1077.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: Macmillan.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of educational opportunity.* Washington, DC: U. S. Government Printing Office.

Comrey, A. L., & Montag, 1. (1982). Comparison of factor analytic results with two choice and seven choice personality item formats. *Applied Psychological Measurement*, *6*, 285-289.

Constantinople, A. (1973). Masculinity-femininity: An exception to a famous dictum?

*Psychological Bulletin*, *80*, 389-407.

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire.* Newbury Park, CA: Sage.

Converse, J. M., & Schuman, H. (1984). The manner of inquiry: An analysis of survey questions form across organizations and over time. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 283-316). New York: Russell Sage Foundation.

Converse, P. E. (1970). Attitudes and non-attitudes: Continuation of a dialogue. In E. R. Tufte (Ed.), *The quantitative analysis of social problems* (pp. 168-189). Reading, MA: Addison-Wesley.

Coombs, C. H. (1950). The concept of reliability and homogeneity. *Educational and Psychological Measurement*, *10*, 43-56.

Cook, T. D., & Campbell, D. T. (1979) *Quasiexperimentation: Design & analysis issues for field settings.* Chicago: Rand McNally.

Coombs, C. H. (1953). Theory and methods of social measurement. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 471-535). New York: Dryden.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Coombs, C. H., & Coombs, L. C. (1976). "Don't know": Item ambiguity or respondent uncertainty. *Public Opinion Quarterly*, *40*, 497-514.

Costner, H. L. (1969). Theory, deduction, and rules of correspondence. *American Journal of Sociology*, *75*, 245-263.

Costner, H. L. (1971). Utilizing causal models to discover flaws in experiments. *Sociometry*, *34*, 398-410.

Couch, A., & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, *60*, 151 - 174.

Cox, E. P. Ⅲ. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, *17*, 407-422.

Coxhead, P., & Bynner, J. M. (1981). Factor analysis of semantic differential data. *Quality and Quantity*, *15*, 553-567.

Coxon, A. P. M. (1982). *The user's guide to multidimensional scaling.* Exeter, NH: Heinemann Educational Books.

Crandall, R. (1973). The measurement of selfesteem and related constructs. In J. P. Robinson & P. R. Shaver (Eds.), *Measures of social psychological attitudes* (rev. ed., pp. 45-167). Ann Arbor, MI: Institute for Social Research.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Cronbach, L. J. (1946). Response set and test validity. *Educational and Psychological Measurement*, *6*, 475-494.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper

& Row.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. I. (1980). Selection theory for a political world. *Public Personnel Management*, *9*, 37-50.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-692). Washington, DC: American Council on Education.

Dawes, R. M. (1972). *Fundamentals of attitude measurement.* New York: Wiley.

Dawes, R. M., & Smith, T. L. (1985). Attitude and opinion measurement. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, 3rd ed., pp. 509-566). New York: Random House.

DeLamater, J. (1982). Response-effects of question content. In W. Dijkstra & J. van der Zouwen (Eds.), *Response behaviour in the survey-interview* (pp. 13-48). London: Academic Press.

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods* (2nd ed.). New York: McGraw-Hill.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement*, *44*, 61-66.

Doby, J. T. (1967). Explanation and prediction. In J. T. Doby (Ed.), *An introduction to social research* (2nd ed., pp. 50-62). New York: Appleton-Century-Crofts.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin*, *86*, 335-337.

Duncan, O. D. (1978). Multiway contingency analysis. *Contemporary Sociology*, *7*, 403-405.

Duncan, O. D. (1984). *Notes on social measurement: Historical and critical.* New York: Russell Sage Foundation.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology*, *30*, 477-525.

Dunn-Rankin, P. (1983). *Scaling methods.* Hillsdale, NJ: Erlbaum.

Edwards, A. L. (1957a). *The social desirability variable in personality assessment.* New York: Dryden.

Edwards, A. L. (1957b). *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts.

Edwards, A. L. (1967a). The social desirability variable: A broad statement. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 32-47). Chicago: Aldine.

Edwards, A. L. (1967b). The social desirability variable: A review of the evidence. In 1. A. Berg (Ed.), *Response set in personality assessment* (pp. 48-70). Chicago: Aldine.

Ehrlich, H, J., & Rinehart, J. W. (1965). A brief report on the methodology of stereotype research. *Social Forces*, *43*, 564-575.

Einhorn, H. J., & Bass, A. R. (1971). Methodological considerations relevant to discrimination in employment testing. *Psychological Bulletin*, *75*, 261-269.

Elliott, W. J. (1983). Ear lobe crease and coronary artery disease: 1000 patients and review of the literature. *American Journal of Medicine*, *75*, 1024-1032.

Ellis, R. J, (1988). Self-monitoring and leadership emergence in groups. *Personality and Social Psychology Bulletin*, *14*, 681-693.

Elmore, P. B., & LaPointe, K. A. (1975). Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. *Journal of Educational Psychology*, *67*, 368-374.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, *43*, 38,290-38,315.

Erdos, P. L. (1970). *Professional mail surveys.* New York: McGraw-Hill.

Erickson, F. (1986). Qualitative methods in research on teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 119-161). New York: Macmillan.

Evertson, C. M., & Green, J. L. (1986). Observation as inquiry and method. In M. C Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 162-213). New York: Macmillan.

Fassnacht, G. (1982). *Theory and practice of observing behaviour* (C. Bryant, Trans.). New York: Academic Press.

Feigl, H., & Brodbeck, M. (Eds.). (1953). *Readings in the philosophy of science.* New York: Appleton-Century-Crofts.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp, 105-146). New York: Macmillan.

Fincher, C. (1975). Differential validity and test bias. *Personnel Psychology*, *28*, 481-500.

Fishbein, M. (Ed.). (1967). *Readings in attitude theory and measurement.* New York: Wiley.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). New York: Hafner.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*, 329-344.

Fitzpatrick, A. R. (1983). The meaning of content validity. *Applied Psychological Measurement*, *7*, 3-13.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist*, *33*, 671-679.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

Fleiss, J. L., & Shrout, P. E. (1977). The effects of measurement errors on some multivariate procedures. *American Journal of Public Health*, *67*, 1188-1191.

Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation procedures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (2nd ed., pp. 253-324). New York: Wiley.

Fowler, F. J. Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error.* Newbury Park, CA: Sage.

Free, L. A., & Cantril, H. (1967). *The political beliefs of Americans: A study of public opinion.* New Brunswick, NJ: Rutgers University Press.

Freedman, D. A. (1987b). A rejoinder on models, metaphors, and fables. *Journal of Educational Statistics*, *12*, 206-223.

Frey, J. H. (1989). *Survey research by telephone* (2nd ed.). Newbury Park, CA: Sage.

Frick, R., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, *48*, 157-184.

Furnham, A. (1984). Work values and beliefs in Britain. *Journal of Occupational Behaviour*, *5*, 281-291.

Gable, R. K. (1986). *Instrument development in the affective domain.* Boston: Kluwer-Nijhoff.

Gage, N. L., Leavitt, G. S., & Stone, G. C. (1957). The psychological meaning of acquiescence set for authoritarianism. *Journal of Abnormal and Social Psychology*, *55*, 98-103,

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*, 564-567.

Galton, F. (1889). *Natural inheritance.* London: Macmfillan.

Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, *45*, 43-57.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review*, *67*, 343-352.

Garner, W. R., Hake, H. W., & Erikson, C. W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*, 149-159.

Geertz, C. (1973). Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures. Selected essays by Clifford Geertz* (pp. 3-30). New York: Basic Books.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Goldstein, T. (1978, February 19). Measuring competence: Debating an indefinable. *The New York Times*, p. E7.

Goocher, B. E. (1965). Effects of attitude and experience on the selection of frequency of adverbs. *Journal of Verbal Learning and Verbal Behavior*, *4*, 193-195.

Goocher, B. E. (1969). More about often. *American Psychologist*, *24*,608-609.

Gorden, R. L. (1975). *Interviewing: Strategy, techniques, and tactics* (rev. ed.). Homewood, IL: Dorsey Press.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Green, B. F. (1981a). A primer of testing. *American Psychologist*, *36*, 1001-1011.

Green, B. F. (Ed.). (1981b). *Issues in testing: Coaching, disclosure, and ethnic bias.* San Francisco: Jossey-Bass.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery—How many scales and response categories to use? *Journal of Marketing*, *34*, 33-39.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827-838.

Greenwald, A. G. (1968). On defining attitude and attitude theory. In A. G. Greenwald, T. C. Brock, & T. M. Ostrom (Eds.), *Psychological foundations of attitudes* (pp. 361-388). New York: Academic Press.

Groves, R. M., & Kahn, R. L. (1979). *Surveys by telephone: A national comparison with personal interviews.* New York: Academic Press.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guion, R. M. (1977). Content validity—The source of my discontent. *Applied Psychological Measurement*, *1*, 1-10.

Guion, R. M. (1978). "Content validity" in moderation. *Personnel Psychology*, *31*, 205-213.

Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, *11*, 385-398.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hagenaars, J. A., & Heinen. T. G. (1982). Effects of role-independent interviewer characteristics on responses. In W. Dijkstra & J. van der Zouwen (Eds.), *Response behaviour in the survey-interview* (pp. 91-130). London: Academic Press.

Hakel, M. D. (1968). How often is often? *American Psychologist*, *23*, 533-534.

Hartley, E. L. (1967). Attitude research and the jangle fallacy. In C. W. Sherif & M. Sherif (Eds.), *Attitude, ego-involvement, and change* (pp. 88-104). New York: Wiley.

Hauser, R. M. (1972). Disaggregating a socialpsychological model of educational attainment. *Social Science Research*, *1*, 159-188.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. In H. L. Costner (Ed.), *Sociological methodology 1971* (pp. 81-117). San Francisco: Jossey-Bass.

Hays, W. L. (1988) *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Hechinger, F. M. (1973, May 27). Home is crucial factor. *The New York Times*, p. E9.

Heise, D. R. (1969b). Some methodological issues in semantic differential research. *Psychological Bulletin*, *72*, 406-422.

Heise, D. R. (1970). The semantic differential and attitude research. In G. F. Summers (Ed.), *Attitude measurement* (pp. 235-253). Chicago: Rand McNally.

Heise, D. R. (1974). Some issues in sociological measurement. In H. L. Costner (Ed.), *Sociological methodology 1973-1974* (pp. 1-16). San Francisco: Jossey-Bass.

Hempel. C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science.* New York: Free Press.

Herbert, J., & Attridge, C. (1975). A guide for developers and users of observation systems and manuals. *American Educational Research Journal*, *12*, 1-20.

Hogarth, R. M. (Edl.). (1982). *Question framing and response frequency.* San Francisco: Jossey-Bass.

Huitema, B. E. (1980). *The analysis of covariance and alternatives.* New York: Wiley.

Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of *test bias*. *Psychological Bulletin*, *83*, 1053-1071.

Hunter, J. S. (1980). The national system of scientific measurement. *Science*, *210*, 869-874.

Hyman, H. H., Cobb, W. J., Feldman, J. J., Hart, C. W., & Stember, C. H. (1954). *lnterviewing in social research.* Chicago: University of Chicago Press.

Jackson, D. J. (1967). Acquiescence response styles: Problems of identification and control. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 71-114). Chicago: Aldine.

Jackson, D. N. (1969). Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin*, *72*, 30-49.

Jacobson, A. L. (1973). Some theoretical and methodological considerations for measuring intrasocietal conflict. *Sociological Methods & Research*, *1*,439-461.

Jacoby, J. (1978). Consumer research: How valid and useful are all our consumer behavior research findings? A state of the art review. *Journal of Marketing*, *42*, 87-96.

Jacoby, J., & Matell, M. S. (1971). Three-point Likert scales are good enough. *Journal of Marketing Research*, *8*, 495-500.

Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, *10*, 93-98.

Johnson, H. G. (1950). Test reliability and correction for attenuation. *Psychometrika*, *15*, 115-119.

Jöreskog, K. G. (1971b). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*, 109-133.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*, 631-639.

Kahn, R. L., & Cannell, C. F. (1957). *The dynamics of interviewing: Theory, technique, and cases.* New York: Wiley.

Kahneman, D. (1963). The semantic differential and the structure of inferences among attributes. *American Journal of Psychology*, 76, 554-567.

Kaiser, H. F. (1960b). Review of *Measurement and statistics: A basic text emphasizing behavioral science applications.* *Psychometrika*, 25, 411-413.

Kalleberg, A. L., & Kluegel, J. R. (1975). Analysis of the multitrait-multimethod: Some limitations and alternatives. *Journal of Applied Psychology*, 60, 1-9.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science.* San Francisco: Chandler.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, *75*, 34-49.

Kazdin, A. E. (1977). Artifact, bias, and com-plexity of assessment: The ABCs of reliability. *Journal of Applied Behavior Analysis*, *10*, 141-150.

Kelley, T. L. (1927). *Interpretation of educational measurements.* Yonkers-on-Hudson, NY: World Book.

Kelly, J. A., Caudill, M. S., Hathorn, S., & O'Brien, C. G. (1977). Socially undesirable sex-correlated characteristics: Implications for androgyny and adjustment. *Journal of Consulting and Clinical Psychology*, *45*, 1185-1186.

Kerlinger, F. N. (1984). *Liberalism and conservatism: The nature and structure of social attitudes.* Hillsdale, NJ: Erlbaum.

Kirscht, J. P., & Dillehay, R. C. (1967). *Dimensions of authoritarianism: A review of research & theory.* Lexington, KY: University of Kentucky Press.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement*, *25*, 987-995.

Kruskal, J. B., & Wish, M. (1978). *Multidimen-sional scaling.* Newbury Park, CA: Sage.

Kubiniec, C. M., & Farr, S. D. (1971). Concept-scale and concept-comPonent interaction in the semantic differential. *Psychological Reports*, *28*, 531-541.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces*, *46*, 151-160.

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review*, *35*,515-524.

Labovitz, S. (1972). Statistical usage in sociology: Sacred cows and ritual. *Sociological Methods & Research*, *1*, 13-37.

Landy, F. J., & Farr, J. L. (1980). Performance ratings. *Psychological Bulletin*, *87*, 72-107.

Larsen, R. J., & Seidman, E. (1986). Gender schema theory and sex role inventories: Some conceptual and psychometric considerations. *Journal of Personality and Social Psychology*, *50*, 205-211.

Lavrakas, P. J. (1987). *Telephone survey methods: Sampling, selection, and supervision.* Newbury Park, CA: Sage.

Lawler, E. E. (1966). Ability as a moderator of the relationship between job attitudes and job performance. *Personnel Psychology*, *19*, 153-164.

Lawler, E. E. (1967). The multitrait-multirater approach to measuring managerial job performance. *Journal of Applied Psychology*, *51*, 369-381.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*, 31-43.

Leamer, E. E. (1988). Discussion. In C. C. Clogg (Ed.), *Sociological methodology 1988* (pp. 485-493). Washington, DC: American Sociological Association.

Lemon, N. (1973). *Attitudes and their measurement.* New York: Wiley.

Lieberson, S. (1985). *Making it count: The improvement of social research and theory.* Berkeley, CA: University of California Peass.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 44-53.

Lin, N. (1976). *Foundations of social research.* New York: McGraw-Hill.

Linn, R. L. (1968). Range of restriction problems in the use of self-selected groups for test validation. *Psychological Bulletin*, *69*, 69-73.

Linn, R. L. (1983a). Person selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, *20*, 1-15.

Linn, R. L. (1983b). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principals* [*sic*] *of modern psychological measurement. A festschrift for Frederic M. Lord* (pp. 27-40). Hillsdale, NJ: Erlbaum.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, *21*, 33-47.

Linn, R. L., & Dunbar, S. B. (1982). Predictive validity of admissions measures: Corrections for selection on several variables. *Journal of College Student Personnel*, *23*, 222-226.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A monte carlo approach. *Journal of Applied Psychology*, *60*, 10-13.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions.* Newbury Park, CA: Sage.

Loeber, R., & Dishion, T. (1983). Early predictors of male delinquency: A review. *Paychological Bulletin*, *94*, 68-99.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, *45*, 507-529.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694 (Monograph Supplement 9).

Lofland, J. (1971). *Analyzing social settings: A guide to qualitative observation and analysis.* Belmont, CA: Wadsworth.

Long, J. S. (1983a). *Confirmatory factor analysis.* Newbury Park, CA: Sage.

Long, V. O. (1989). Relation of masculinity to self-esteem and self-acceptance in male professionals, college students, and clients. *Journal of Counseling Psychology*, *36*, 84-87.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750-751.

Lord, F. M. (1954). Further comment on "Football Numbers." *American Psychologist*, *9*, 264-265.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mackay, A. L. (1977). *The harvest of a quiet eye: A selection of scientific quotations.* Bristol, UK: Institute of Physics.

Maguire, T. O. (1973). Semantic differential methodology for the structuring of attitudes.

*American Educational Research Journal*, *10*, 295-306.

Mann, I. T, Phillips, J. L. , & Thompson, E. G. (1979). An examination of methodological is sues relevant to the use and interpretation of the semantic differential. *Applied Psychological Measurement*, *3*, 213-229.

Maranell, G. M. (1974a). Introduction. In G. M. Maranell (Ed.), *Scaling: A sourcebook for behavioral scientists* (pp. xi-xix). Chicago: Aldine.

Maranell, G. M. (Ed). (1974b). *Scaling: A sourcebook for behavioral scientists.* Chicago: Aldine.

Margenau, H. (1959). Philosophical problems concerning the meaning of measurement in physics. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 163-176). New York: Wiley.

Markus, H. , & Zajonc, R. B. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. I, 3rd ed. , pp. 137-230). New York: Random House.

Marradi, A. (1981). Factor analysis as an aid in the formulation and refinement of empirically useful concepts. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 11-49). Newbury Park, CA: Sage.

Marsh, C. (1982). *The survey method: The contribution of surveys to sociological explanation.* London: Allen & Unwin.

Martin, E. (1983). Surveys as social indicators: Problems in monitoring trends. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 677-743). New York: Academic Press.

Matell, M. S. , & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study Ⅰ: Reliability and validity. *Educational and Psychological Measurement*, *31*, 657-674.

Matell, M. S. , & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale proper-ties. *Journal of Applied Psychology*, *56*,506-509.

Mayerberg, C. K. , & Bean, A. G. (1978). Two types of factors in the analysis of semantic differential attitude data. *Applied Psychological Measurement*, *2*, 469-480.

McGaw, B. , Wardrop, J. L. , & Bunda, M. A, (1972). Classroom observation schemes: Where are the errors? *American Educational Research Journal*, *9*, 13-27.

McHugh, R. B. (1957). The interval estimation of a true score. *Psychological Bulletin*, *54*, 73-74.

Mclver, J. P. , & Carmines, E. G. (1981). *Unidimensional scaling.* Newbury Park, CA:Sage.

McKelvie, S. G. (1978). Graphic rating scales: How many categories? *British Journal of Psychology*, *69*, 185-202.

McNeil, B. J. , Pauker, S. G. , Sox, H. C. , & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *The New England Journal of Medicine*, *306*, 1259-1262.

Medley, D. M. (1982). Systematic observation. In H. E. Mitzel (Ed,), *Encyclopedia of educational research* (5th ed. , pp. 1841-1851). New York: Free Press.

Medley, D. M. , & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 247-328). Chicago: Rand McNally.

Meehl, P. E. , & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194-216.

Messick, S. J. (1967). The psychology of acquiescence: An interpretation of research evidence. In I. A. Berg (Ed.), *Response set in personality assessment* (pp. 115-145). Chicago: Aldine.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, *89*, 575-588.

Milavsky, J. R. , Kessler, R. C. , Stipp, H. H. , & Rubens, W. S. (1984). *A comment* by J. Ronald Milavsky, Ronald C. Kessler, Horst H. Stipp, and William S. Rubens. *Journal of Communication*, *34*, 182-187.

Miller, D. C. (1983). *Handbook of research design and social measurement* (4th ed.). New York: Longman.

Miller, P. V. , & Cannell, C. F. (1988). Interviews in sample surveys. In J. P. Keeves (Ed.), *Educational research*, *methodology*, *and measurement*: *An international handbook* (pp. 457-465). New York: Pergamon.

Miron, M. S. (1972). Universal semantic differential shell game. *Journal of Personality and Social Psychology*, *24*, 313-320.

Miron, M. S. , & Osgood, C. E. (1966). Language behavior: The multivariate structure of qualification. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 790-819). Chicago: Rand McNally.

Mischel, W. , Zeiss, R. , & Zeiss, A. (1974). Internal-external control and persistence: Validation and implications of the Stanford preschool internal-external scale. *Journal of Personality and Social Psychology*, *29*, 265-278.

Mishler, E. G. (1986). *Research interviewing*: *Context and narrative.* Cambridge, MA: Harvard University Press.

Mitchell, S. K. (1979). lnterobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86*, 376-390.

Molenaar, N. J. (1982). Response-effects of "for-mal" characteristics of questions. In W. Dijkstra & J. van der Zouwen (Eds.), *Response behaviour in the survey-interview* (pp. 49-89). London: Academic Press.

Montaigne. (1965). *The complete essays of Montaigne* (D. M. Frame, Trans.). Stanford:

Stanford University Press.

Nagel, E. (1931). Measurement. *Erkenntnis*, 2, 313-333.

Northrop, F. S. C. (1947). *The logic of the sciences and the humanities.* New York: Macmillan.

Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1-13.

Nunnally, J. (1967). *Psychometric theory.* New York: McGraw-Hill.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Oppenheim, A. N. (1966). *Questionnaire design and attitude measurement.* New York: Basic Books.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*, 197-237.

Osgood, C. E., Suci, G. J., & Tennenbaum, P. H. (1957). *The measurement of meaning.* Urbana, IL: University of Illinois.

Ostrom, T. M. (1969). The relationship between the affective, behavioral, and cognitive components of attitude. *Journal of Experimental Social Psychology*, *5*, 12-30.

Page, M. M. (Ed.). (1980). *Beliefs, attitudes, and values: 1979 Nebraska Symposium on Motivation.* Lincoln, NE: University of Nebraska Press.

Paulhus, D. L., & Martin, C. L. (1988). Functional flexibility: A new conception of interpersonal flexibility. *Journal of Personality and Social Psychology*, *55*, 88-101.

Payne, S. L. (1951). *The art of asking questions.* Princeton: Princeton University Press.

Peaker, G. F. (1975). *An empirical study of education in twenty-one countries: A technical report.* New York: Wiley.

Pedhazur, E. J. (1978). Wilson-Patterson Attitude Inventory. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (Vol. 1, pp. 1150-1152). Highland Park, NJ: Gryphon Press.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston.

Pedhazur, E. J., & Tetenbaum, T. J. (1979). Bem sex role inventory: A theoretical and methodological critique. *Journal of Personality and Social Psychology*, *37*, 996-1016.

Petersen, N. S. (1980). Bias in the selectionrule-Bias in the test. In L. J. Th. van der Kemp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 103-122). New York: Wiley.

Petersen, N. S., & Novick, M. R. (1976). An evaluation of some models of culture-fair selection. *Journal of Educational Measurement*, *13*, 3-29.

Pezzullo, T. R., & Brittingham, B. E. (Eds.). (1979). *Salary equity: Detecting sex bias in salaries among college and university professors.* Lexington, MA: Heath.

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach.* Oxford: ClarendonPress.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, *38*, 513-532.

Reiser, M., Wallace, M., & Schuessler, K. (1986). Direction-of-wording effects in dichotomous social life feeling items. In N. B. Tuma (Ed.), *Sociological methodology 1986* (pp. 1-25). San Francisco: Jossey-Bass.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178-208). New York: Wiley.

Reynolds, C. R., & Brown, R. T. (Eds.). (1984). *Perspectives on bias in mental testing.* New York: Plenum.

Richards, I. A. (1926). *Principles of literary criticism* (2nd ed.). London: Routledge & Kegan Paul.

Roberts, A. O. H. (1980). Regression toward the mean and the regression-effect bias. In G. Echternacht (Ed.), *Measurement aspects of Titie I evaluations* (pp. 59-82). San Francisco: Jossey-Bass.

Robinson, W. S. (1957). The statistical measurement of agreement. *American Sociological Review*, *22*, 17-25.

Rossiter, C. (1968). Conservatism. In D. L. Sills (Ed.), *International encyclopedia of the social sciences* (Vol. 3, pp. 290-295). New York: Macmillan.

Rowley, G. L. (1976). The reliability of observational measures. *American Educational Research Journal*, *13*, 51-59.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, IL: Dorsey Press.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, *88*, 413-428.

Sackett, G. P. (Ed.). (1978). *Observing behavior Volume II: Data collection and analysis methods.* Baltimore: University Park Press.

Sandelands, L. E., & Calder, B. J. (1984). Referencing and bias in social interaction. *Journal of Personality and Social Psychology*, *46*,755-762.

Sanford, N. (1973). Authoritarian personality in contemporary perspective. In J. N. Knutson (Ed.), *Handbook of political psychology* (pp. 139-170). San Francisco: Jossey-Bass.

Schmitt, N., Coyle, B. W., & Saari, B. B. (1977). A review and critique of analyses of multitrait-multimethod matrices. *Multivariate Behavioral Research*, *12*,447-478.

Schmitt, N., & Stults, D. M. (1986). Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, *10*, 1-22.

Schneider, B. (1970). Relationships between various criteria of leadership in small groups. *Journal of Social Psychology*, *82*, 253-261.

Schneider, D. J., Hastorf, A. H., & Ellsworth, P. C. (1979). *Person perception* (2nd ed.). Reading, MA Addison-Wesley.

Schuman, H., & Kalton, G. (1985). Survey methods. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, 3rd ed., pp. 635-698). New York: Random House.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on*

*question form, wording, and context.* New York: Academic Press.

Scott, W. A. (1960). Measures of test homogeneity. *Educational and Psychological Measurement*, *20*, 751-757.

Scriven, M. (1959). Explanation and prediction in evolutionary theory. *Science*, *130*, 447-482.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement*, *23*, 153-158.

Shavelson, R. J., & Bolus, R. (1982). Self-concept: The interplay of theory and methods. *Journal of Educational Psychology*, *74*, 3-17.

Shavelson, R. J., Webb, N. M., & Burstein, L. (1986). Measurement of teaching. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 50-91). New York: Macmillan.

Sheatsley, P. B. (1951). The art of interviewing and a guide to interviewer selection and training. In M. Jahoda, M. Deutsch, & S. W. Cook (Eds.), *Research methods in social relations* (pp, 463-492). New York: Dryden.

Sbeatsley, P. B. (1983). Questionnaire construction and item writing. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 195-230). New York: Academic.

Simpson, R. H. (1944). The specific meaning of certain terms indicating differing degrees of frequency. *The Quarterly Journal of Speech*, *30*, 328-330.

Smith, T. W. (1984). Nonattitudes: A review and evaluation. In C. F. Turner & E. Martin (Eds.), *Surveying subjective phenomena* (Vol. 2, pp. 215-255). New York: Russell Sage Foundation.

Snider, J. G., & Osgood, C. E. (Eds.). (1969). *Semantic differential technique.* Chicago: Aldine.

Sockloff, A. L. (Ed.) (n. d.). *Proceedings. The First Invitational Conference on Faculty Effectiveness as Evaluated by Students.* Philadelphia: Measurement and Research Center, Temple University.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Stamp, J. (1929). *Some economic factors in modern life.* London: King.

Stanley, J. C. (1971). Reliability. In R. L. Thorn-dike (Ed.), *Educational measurement* (2nd ed., pp. 356-442). Washington, DC: American Council on Education.

Stanley, J. C., & Wang, M. D (1970). Weighting test items and test-item options, an overview of the analytical and empirical literature. *Educational and Psychological Measurement*, *30*, 21-35.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

Stevens, S. S. (1959). Measurement, psychophysics, and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 18-63). New York: Wiley.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*, 849-856.

Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin*, *105*, 147-155.

Sudman, S., & Bradburn, N. M. (1974). *Response effects in surveys: A review and synthesis.* Chicago: Aldine.

Sudman, S., & Bradburn, N. M. (1982). *Asking questions.* San Francisco: Jossey-Bass.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, *23*, 565-578.

Tenopyr, M. L. (1977). Content—construct confusion. *Personnel Psychology*, *30*, 47-54.

Terwilliger, J. S., & Lele, K. (1979). Some relationships among internal consistency, reproducibility, and homogeneity. *Journal of Educational Measurement*, *16*, 101-108.

Tesser, A., & Krauss, H. (1976). On validating a relationship between constructs. *Educational and Psychological Measurement*, *36*, 111-121.

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques.* New York: Wiley.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560-620). Washington, DC: American Council on Education.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin.

Thorndike, R. L., & Hagen, E. P. (1977). *Measurement and evaluation in psychology and education* (4th ed.). New York: Wiley.

Thurstone, L. L. (1947). *Multiple-factor analysis.* Chicago: Chicago University Press.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology 22*, 358-376.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Towstopiat, O. (1984). A review of reliability procedures for measuring observer agreement. *Contemporary Educational Psychology*, *9*, 333-352.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. *Psychological Bulletin 54*, 229-249.

Turner, C. F., & Martin, E. (Eds.). (1984). *Surveying subjective phenomena* (Vols. 1-2). New York: Russell Sage Foundation.

van der Ven, A. H. G. S. (1980). *Introduction to scaling.* New York: Wiley.

Vernon, P. E. (1972). The distinctiveness of field independence. *Journal of Personality*, *40*, 366-391.

Wagner, R. F., Reinfeld, H. B., Wagner, K. D., Gambino, A. T., Falco, T. A., Sokol, J. A., Katz, S., & Zeldis, S. M. (1984). Ear-canal hair and the ear-lobe crease as predictors for coronary-artery disease. *The New England Journal of Medicine*, *311*, 1317-1318.

Wainer, H. (1987). *The first four millennia of mental testing: From ancient China to the computer*

*age.* Princeton, NJ: Educational Testing Service.

Wainer, H., & Braun, H. 1. (Eds.). (1988). *Test validity.* Hillsdale, NJ: Erlbaum.

Wallace, S. R. (1965). Criteria for what? *American Psychologist*, *20*, 411-417.

Wallis, W. A., & Roberts, H. V. (1956). *Statistics: A new approach.* New York: Free Press.

Wanous, J. P., & Lawler, E. E. (1972). Measurement and meaning of job satisfaction. *Journal of Applied Psychology*, *56*, 95-105.

Warr, P. B., & Knapper, C. (1968). *The perception of people and events.* London: Wiley.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, *21*, 81-90.

Weick, K. E. (1968). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, 2nd ed., pp. 357-451). Reading, MA: Addison-Wesley.

Weick, K. E. (1985). Systematic observational methods. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, 3rd ed., pp. 567-634). New York: Random House.

Weinberg, E. (1983). Data collection: Planning and management. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), *Handbook of survey research* (pp. 329-358). New York: Academic.

Weiss, D. J., & Davison, M. L. (1981). Test theory and methods. *Annual Review of Psychology*, *32*, 629-658.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Wilson, G. D. (1975). *Manual for the Wilson-Patterson Attitude Inventory (WPAI).* London: NFER.

Winne, P. H., & Belfry, M. J. (1982). Interpretive problems when correcting for attenuation. *Journal of Educational Measurement*, *19*, 125-134.

Witkin, H. A., Dyk, R. B., Faterson, H. F., Goodenough, D. R., & Karp, S. A. (1962). *Psychological differentiation: Studies of development.* New York: Wiley.

Wolins, L. (1982). *Research mistakes in the social and behavioral sciences.* Ames, IA: Iowa State University Press.

Wolinsky, J. (1983, July). Masters, Johnson respond to criticism. *APA Monitor*, p. 2.

Wylie, R. C. (1974). *The self-concept: A review of methodological considerations and measuring instruments* (Vol. 1, 2nd ed.). Lincoln, NE: University of Nebraska Press.

Zeller, R. A., & Carmines, E. G. (1980). *Measurement in the social sciences: The link between theory and data.* Cambridge: Cambridge University Press.

[General Information]
书名=定量研究基础：测量篇
页数=205
SS号=13236466